J Stat Phys (2013) 153:442–459
DOI 10.1007/s10955-013-0833-6

Predictive Information in a Nonequilibrium Critical Model

Martin Tchernookov · Ilya Nemenman

Received: 17 December 2012 / Accepted: 16 August 2013 / Published online: 13 September 2013

Abstract We propose predictive information, that is, information between a long past of duration T and the entire infinitely long future of a time series, as a general order parameter to study phase transitions in physical systems independently of the underlying dynamics. It can be used, in particular, to study nonequilibrium transitions and other exotic transitions, where a simpler order parameter cannot be identified using traditional symmetry arguments. As an example, we calculate predictive information for a stochastic nonequilibrium dynamics problem that forms an absorbing state under a continuous change of a parameter. The information at the transition point diverges as $\propto \log T$, and we calculate the expression for a smooth crossover to $\propto T^0$ away from the transition.

Keywords Phase transitions · Information theory · Subextensive scaling

1 Introduction

The theory of critical phenomena and the emergent notion of universality was one of the singular developments of physics in the twentieth century. With a known order parameter and symmetries of the problem, calculation of long-range, measurable behaviors of equilibrium physical quantities becomes a rather straightforward task. The success has turned out to be hard to replicate for nonequilibrium systems and systems where symmetry properties are similar in the phases on both sides of the transition [1]. Here it is often unclear which quantity can serve as a good order parameter, and the developed theoretical machinery does not apply. Where progress has been made, order parameters have been very specific, making it difficult to identify universal properties. For example, in reaction-diffusion problems

M. Tchernookov (✉)
Department of Physics, Emory University, Atlanta, GA 30322, USA
e-mail: mtchern@emory.edu

I. Nemenman
Departments of Physics and Biology and Computational and Life Sciences Initiative, Emory University,
Atlanta, GA 30322, USA
e-mail: ilya.nemenman@emory.edu

with absorption, one commonly uses linear superposition of particle concentrations as order parameters [2, 3], while particle current is a better choice for jamming problems [4]. Further, the order parameters often have nontrivial relations to easily observable quantities. For example, phase transitions in some systems with dynamic heterogeneities often must be described with four-point correlation functions of particle densities [5], or a multitude of correlation functions [6, 7]. Similarly, dynamical phase transition require one to study the space of trajectories instead of the state space [8]. The latter approach, known as the method of large deviations, can be modified to describe glassy systems [9–11].

Whatever the choice, the order parameter is a statistics averaged over a distribution of microscopic states. A continuous or discontinuous change in its value at a transition indicates a similar change in the underlying probability distribution. Therefore, it is natural to shift attention to the distribution itself. For example, one can analyze the spectrum of the operator that controls the evolution of the probability distribution [12], or work directly with the stationary state [13]. In this paper, we focus on the distribution of a nonequilibrium system as it converges to the steady state.

Intuitively, different phases (often with different symmetries) manifest themselves by changes in our ability to use local experimental measurements for long-range predictions. For example, nonzero magnetization in an Ising magnet allows us to predict with some certainty orientation of far away spins based on the value of the spin at the origin. Similarly, different crystalline phases of solids have different density autocorrelation functions, and hence existence of an atom at the origin translates into different predictions about the presence of an atom a certain distance away. Then instead of a specific statistics characterizing the predictability, namely the order parameter, it might be useful to study one's ability to use local measurements to predict states of the rest of the system *directly*.

This prediction ability is naturally quantified using the language of Shannon's information theory [14]. In previous work, we have termed it the *predictive information* [15, 16]. Briefly, in information theory, the total uncertainty in a system specified by a state $\mathbf{x} \in X$, $\dim \mathbf{x} = N$, is measured by the (differential) entropy,

$$S[X] = -\int d^N x\, P(\mathbf{x}) \log_2 P(\mathbf{x}). \tag{1}$$

Then observing a state of another variable $\mathbf{y} \in Y$, $\dim y = M$, may reduce the uncertainty about $\mathbf{x}$, and hence provide the information about it

$$\begin{aligned} I[X;Y] &= S[X] - \langle S[X|Y] \rangle_Y = \int d^N x\, d^M y\, P(\mathbf{x},\mathbf{y}) \log_2 \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{x})P(\mathbf{y})} \\ &= \left\langle \log_2 \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{x})P(\mathbf{y})} \right\rangle_{X,Y} = I[Y;X]. \end{aligned} \tag{2}$$

Importantly, $I[X;Y]$ depends on the entire probability distribution $P(\mathbf{x},\mathbf{y})$, but not just on its specific statistics, and it is zero iff X and Y are statistically independent.

One can consider X and Y to be states of a physical process, such that X are the measured quantities, and Y are the quantities that one wants to predict [15]. For example, X can be the state of spins on one segment of an Ising chain, and Y be the state of spins far away. Similarly, for time series and for non-equilibrium processes, X can be the past of the process of duration N, and Y part of its future of duration M. Then the information becomes the *predictive information*:

$$I_{\text{pred}}(N, M) = I[X;Y]. \tag{3}$$

Since the quantification of the intrinsic state of the system should not depend on which specific set of variables Y one wants to predict, it makes sense to define predictive information as

$$I_{\text{pred}}[X] \equiv I_{\text{pred}}(N) = \lim_{M\to\infty} I(N, M). \quad (4)$$

That is, one quantifies how much information the local observations X provide about an entire, infinitely large physical system. The predictive information is an averaged quantity. Thus it can depend only on the length N and the initial distribution at the beginning of time series. In our notation, the latter dependence is assumed but not stated. This is because unless there is explicit symmetry breaking, the infrared behavior of the system is not affected by the initial conditions. Therefore, we expect the asymptotic behavior of I_{pred} to be, up to a constant, only a function of N. Indeed, we will demonstrate this in the subsequent Sections.

Predictive information is subextensive, $\lim_{N\to\infty} I_{\text{pred}}(N)/N = 0$, for stationary processes [15], and it is function only of subextensive components of the involved multivariate entropies. It tends to a handful of universal behaviors for large systems, $N \to \infty$, intuitively correlating with the complexity of the underlying physical process. In particular, $\lim_{N\to\infty} I_{\text{pred}}(N) = \text{const}$ indicates an easily predictable deterministic, or a short correlation length probabilistic dynamics ("simple" long range prediction can be perfect, or it is impossible, respectively). For example, below T_c in an Ising ferromagnet, the entire prediction is limited to knowing which of the two states (up or down) the entire system is in, and hence the predictive information is, at most, a bit. Further, $\lim_{N\to\infty} I_{\text{pred}}(N) \propto \log N$ is indicative of a second order equilibrium phase transition (power-law decaying correlations allow for complex, multiscale, partially predictable patterns over very long distances). Finally, $\lim_{N\to\infty} I_{\text{pred}}(N) \sim N^{\alpha}$, $\alpha < 1$, $N \to \infty$, may correspond to more exotic phase transitions with infinite-dimensional order parameters, but this case is not well understood.

The dependence of I_{pred} on the full underlying probability distribution and the relation to phase transitions make it natural to explore I_{pred} as a *general* order parameter, also useable in the nonequilibrium context. In fact, related properties of predictive information and similar quantities has been explored repeatedly in different contexts, such as defining complexity of a time series [15, 17–20], finding meaningful information in data [16, 21–23], or studying quantum entanglement [24], or thermodynamics of prediction [25]. However, we are not aware of calculations of predictive information for nonstationary processes, where $P(\mathbf{x})$ is explicitly or implicitly time dependent. Further, even for equilibrium systems, the transition between $I_{\text{pred}} = \text{const}$ and $I_{\text{pred}} \propto \log N$ in the vicinity of a phase transition has not been studied explicitly.

In this paper, we study predictive information in a context of a simple nonequilibrium, continuous-time Markov process, which ages and develops an absorbing state at a certain critical value of a parameter. This process can be viewed as a toy model, which is likely to possess some features of more complex systems. We calculate the expression for predictive information at the critical point and, for the first time for any system, near the critical point. The calculation reveals the need to modify the definition, Eq. (4), to remove an ultraviolet divergence emerging due to the continuous-time nature of the process. Similar modifications will likely allow extension of predictive information methodology to multidimensional systems. We demonstrate explicitly the logarithmic divergence of I_{pred} at the transition, and we show that the divergent term in the information is insensitive to temporally local, invertible transformations of the state space. This makes predictive information, and specifically its divergent term, a great candidate to characterize nonequilibrium phase transitions.

2 The Model

We consider a Markovian system governed by the following Langevin equation:

$$\partial_t x(t) = -x\left(x^2+\tau\right) + \sqrt{2}\sigma|x|^{\alpha/2}\eta, \tag{5}$$

$$x(t=0) = x_0, \quad \text{sampled from } P(x_0) \equiv P_0, \tag{6}$$

where $\langle\eta(t)\eta(t')\rangle = \delta(t-t')$. We will treat this equation in the Ito sense. Without the noise term, x relaxes from the initial value x_0 to either 0 or $\pm\sqrt{\tau}$, depending on if $\tau > 0$. The transition happens at $\tau = 0$. For large noise near $x = 0$ (that is, small α), x gets kicked out from $x \approx 0$ region, and the system equilibrates. For small noise (large α), a near-deterministic relaxation to the absorbing state at $x = 0$ persists. This is probably the simplest example of nonequilibrium, stochastic relaxation dynamics, and it is a natural point for the first analysis. The model resembles the Simple Exclusion processes and related models describing Kipnis-Marchioro-Presutti transport of energy [26, 27]. However, the similarity is superficial since Simple Exclusion models are conservative and $1+1$ dimensional. Hence they do not develop an absorbing state and exhibit a very different critical point. Instead, we emphasize that, even though our model is one-dimensional, and the noise and the force terms are not explicitly time-dependent, this is a true nonequilibrium phase transition that emerges because the variance of the noise is time-dependent and is allowed to go to zero with x. This reduction of variance is an indication of aging in the model. In turn, this implies that the system is capable of remembering its past, which could be interpreted as the result of long-range effective interactions. This aging, and hence absence of stationarity, means that the predictive information is not necessarily subextensive in this system, making the problem richer.

The model in Eq. (5) is manifestly a toy model, without a direct relation to realistic systems. Nonetheless, some intuition can be obtained since individual terms in Eq. (5) are easy to understand. Specifically, the force is the lowest order expansion of any smooth force around the absorbing point $x = 0$, and hence is quite general. The noise term is also of a general form that would allow for a phase transition to an absorbing state, and similar noises have been discussed in the literature [28]. Intuitively, we can view the model as describing dynamics of magnetization, x, along a line normal to a boundary of an Ising ferromagnet in some number of spatial dimensions. The coordinate is $t = 0$ at the boundary, and increases into the bulk. The deterministic cubic dynamics in Eq. (5) is the usual coarse-grained model of such ferromagnet. In such a model, the variance of the noise increases with x, and α would depend on the overall dimensionality of the problem.

Before we commit ourselves to calculating the predictive information for the dynamics in Eq. (5), it is worthwhile to discuss utility of this exercise for the case when a perfectly good, simple order parameter exists (x itself). To identify and characterize phase transitions in more complicated systems, where choice of the order parameter is nontrivial, one needs to know how such transitions manifest themselves in I_{pred}. This signature is currently unknown. Thus to establish predictive information as a general order parameter, it makes sense first to focus on a system where the transition is identifiable by other means, and then to study I_{pred} near the transition, identifying the sought after behavior. Therefore, straightforward analytical tractability is another argument for the choice of Eq. (5) for the first investigation.

3 Preliminaries

To calculate predictive information, Eq. (3), we discretize the time t, $t_n = n\Delta t$, and $x_n = x(t_n)$. We choose $\Delta t \to 0$, and yet $N\Delta t = T_{\text{p}} \to \infty$, and $M\Delta t = T_{\text{f}} \to \infty$, where p and f stand for *past* and *future*, respectively. Then Eq. (5) is equivalent to the following Markovian dynamics:

$$\begin{aligned} P(x_{n+1}|x_0, x_1, \ldots, x_n) &= P(x_{n+1}|x_n) \\ &= \frac{1}{\sqrt{4\pi\,\Delta t}\sigma x^{\alpha/2}} \exp\left\{-\frac{[x_{n+1} - (x_n - x_n(x_n^2 + \tau)\Delta t)]^2}{4\sigma^2 |x_n|^\alpha \Delta t}\right\}. \end{aligned} \tag{7}$$

To simplify the notation, we define

$$P_{n|n-1} \equiv P(x_n|x_{n-1}), \tag{8}$$

$$P_n \equiv P(x_n) = \int dx_{n-1} P(x_{n-1}) P(x_n|x_{n-1}), \tag{9}$$

$$I_{\mathrm{pred}}(N, M) \equiv I\big[X = \{x_i\}_{i=0}^{N-1}; Y = \{x_i\}_{i=N}^{N+M-1}\big]. \tag{10}$$

Then:

$$\begin{aligned} I_{\mathrm{pred}}(N, M) &= \left\langle \log_2 \frac{P_0 \prod_{n=1}^{N+M-1} P_{n|n-1}}{P_0 \prod_{n=1}^{N-1} P_{n|n-1} P_N \prod_{m=N+1}^{N+M-1} P_{m|m-1}} \right\rangle \\ &= \left\langle \log_2 \frac{P_{N|N-1}}{P_N} \right\rangle = I[x_N; x_{N-1}]. \end{aligned} \tag{11}$$

Not surprisingly for a Markovian process, predictive information is the mutual information between two successive measurements and does not depend on the length of the future sequence, M, so that the limit, Eq. (4), is trivial. However, the information can depend on N since the system is not stationary, and not time-translation invariant. Specifically, for small noise, each subsequent x is more narrowly distributed. This allows the information to increase unboundedly with N, unlike in typical finite-dimensional Markov processes with constant transition probabilities, where I_{pred} is always finite [15]. These considerations also point out that one must take the sequence on N observations starting from exactly the same time when calculating the averages.

Since $x(t)$ is continuous, $x_N \to x_{N-1}$ as $\Delta t \to 0$. The state of the process at the next time step becomes exactly known, and predictive information diverges. This issue always arises in the continuous limit of information quantities, and it is well-known in the large deviation literature in the context of the Kolmogorov-Sinai entropy [29, 30]. However, this is a superficial ultraviolet divergence, due to the infinitesimally small time delay between the past and future sequences, whose mutual information we are calculating. Therefore, it is a boundary effect. Instead, we are interested in studying the infrared behavior. Interestingly, this interfacial effect has been the primary reason behind the inability to apply predictive information ideas to systems in more than one dimension, where the size of the interface diverges with the system size. This makes it difficult to disambiguate divergences in predictive information coming from long-range prediction from those produced by short range interfacial effects.

We thus need to introduce the cutoff scale into the system, at which predictive information is computed, similarly to how one does this in the renormalization group theory. For this, we redefine predictive information as mutual information between the past of duration $T_{\mathrm{p}} = N\Delta t$ and the future of duration $T_{\mathrm{f}} = M\Delta T$, separated by a "scale" gap of duration

$T_\mathrm{s} = L\Delta T$, which remains finite as $\Delta T \to 0$. That is

$$I_\mathrm{pred}(N, M|L) = \left\langle \log_2 \frac{P_0 \prod_{n=1}^{N-1} P_{n|n-1} P_{N+L|N-1} \prod_{m=N+L+1}^{N+L+M-1} P_{m|m-1}}{P_0 \prod_{n=1}^{N-1} P_{n|n-1} P_{N+L} \prod_{m=N+L+1}^{N+L+M-1} P_{m|m-1}} \right\rangle$$

$$= \left\langle \log_2 \frac{P_{N+L|N-1}}{P_{N+L}} \right\rangle = I[x_{N+L}; x_{N-1}]. \tag{12}$$

Here

$$P_{N+L|N-1} = \int \prod_{n=N}^{N+L-1} dx_n \prod_{m=N}^{N+L} P_{m|m-1}. \tag{13}$$

4 Invariance of Predictive Information

From Eq. (12), it is clear that predictive information is invariant under reparameterization of x. This is a desired property for any potential general order parameter, so that one does not need to make a specific choice of parameterization of x to study asymptotic properties of predictive information. However, the states x are pure in the sense that they represent the system at single moments in time. On the other hand, any experimental device measuring $x(t)$ will act as a temporal filter $\mathcal{F}$, so that the measured values will be convolutions of true x's at nearby time points. While in general the filtered data might not have the same predictive properties, it is desirable for the nonequilibrium order parameter to be invariant to a certain class of transformations, namely temporally local invertible filters [15]. In the following, we propose a precise definition of such transformations.

The filter, represented by $\mathcal{F}$, maps the sequences of true states of the system $\{x\}$ into measured data $\{\chi\}$. A filter describing an experimental device has internal degrees of freedom, which influence the measurements. Our notion of a general order parameter refers only to the underlying dynamics and not the details of the experimental procedure. Therefore, we propose an idealized scenario in which we require that the filter does not inject additional information into the dynamics. This means that the extraneous parameters of the mapping $\mathcal{F}$ must be known and the mapping itself must be translationally invariant. In a real-life experiment, this means that we would like to be able to separate the behavior of the observed system from any artifacts associated with the experimental setup. In general terms, such a filter can be represented by a convolution kernel $\mathcal{L}(t-t')$, where all parameters of the function $\mathcal{L}$ are known. Since a convolution mixes the past and the future, the measured data $\{\chi\}$ is no longer Markovian. We would like to preserve the asymptotic behavior of the predictive information, therefore we require that the so-introduced statistical dependences are short lived, i.e. the kernel $\mathcal{L}(t-t')$ is of compact support or decreases with time exponentially or faster. This is our definition of temporal locality. We would like to verify to which extent I_pred calculated for the sequence $\{x\}$ is the same as I_pred calculated for the sequence $\{\chi\}$, if the two sequences are related by such (invertible) local filters.

Since convolutions are reductions in rank, defining invertibility is not trivial and is possible only for infinitely long data sequences. Therefore, we can define invertibility only in the $t \to \infty$ limit. To this end, let $\mathfrak{V} = \bigotimes_n \mathbb{R}^n$ be the space of all temporally discretized, finite length trajectories, that is the space of all n-tuples of x, $n < \infty$. Let $\mathcal{F} : \mathfrak{V} \to \mathfrak{V}$ be a function such that $\mathcal{F}(\mathbb{R}^{N+\nu}) \subset \mathbb{R}^N$. That is, a sequence of N data points is defined from $N+\nu$ points through some filtering procedure. We consider this mapping to be invertible if the Radon-Nikodym derivative over the set $\mathcal{F}^{-1}(\mathbf{x} \in \mathbb{R}^N)$ converges to a delta function for $N \to \infty$. More specifically, the probability of observing a trajectory $\{\chi_i\}_{i=1}^N$ is given by

$$\begin{aligned} P\big(\{\chi_j\}_{i=j}^N\big) &= \int d^{N+\nu}x\, P\big(\{x_j\}_{j=-\nu}^N\big) \prod_{j=1}^N \delta\Big(\chi_j - \sum_k \mathcal{L}(j-k)x_k\Big) \\ &= \int d^{N+\nu}x\, d^N\lambda \times \exp\Bigg[-i\sum_{j=1}^N \lambda_j\Big(\chi_j - \sum_k \mathcal{L}(j-k)x_k\Big) + \ln P\big(\{x_j\}_{j=-\nu}^N\big)\Bigg]. \end{aligned} \tag{14}$$

Thus invertibility requires that the Hessian matrix of the exponent in this equation diverges, defining a dominant stationary solution of the corresponding "action". With this requirement, $\{\chi_i\}$ are simply reparameterizations of $\{x_i\}$, and predictive information is invariant under the change (up to $O(1)$ corrections due to the end points of the sequences). While this requirement is very general, we suspect that, in practice, it will be equivalent to the asymptotic properties of trajectory-averaged quantities, for which there are already well established results [31], and hence the *divergent* component of I_{pred} is invariant under invertible, temporally local filters. We leave exploration of these conditions to future work, instead, we provide here the following simple example.

Let $\mathcal{F}$ be defined through taking the average between adjacent points, i.e. $\mathcal{F}(\{x_n\}_{n=0}^N) = \{(x_n + x_{n-1})/2\}_{n=1}^N$. If the underlying dynamics is purely diffusive, for the probability of observing a sequence of data $\{\chi_i\}_{i=1}^N$ we can write

$$\begin{aligned} P\big(\{\chi\}\big) = \int dk P(x_0 + k)(2\pi\Delta t)^{-N/2} \\ \times \exp\Bigg(-\frac{1}{2\Delta t}\sum_{n=1}^N \big[x_n - x_{n-1} + 2(-1)^n k\big]^2\Bigg) \end{aligned} \tag{15}$$

Here, Δt is the time separation between adjacent trajectory points, and $\{x\}_{n=0}^N$ is the unique solution to $\chi = \mathcal{F}(\mathbf{x})$ with the initial condition $x_0 = x_1$, i.e., $x_0 = x_1 = \chi_1$, $x_n = (-1)^{n-1}\chi_1 + 2\sum_{i=2}^n (-1)^{n-i}\chi_i$. Let $k = k_c + \delta k$ and notice that the expression in the above exponent is quadratic in k. Therefore, we can choose k_c to minimize the sum in Eq. (15):

$$k_c = \frac{\sum_{n=0}^{N-1}(-1)^n(x_{n+1} - x_n)}{2N} \tag{16}$$

Thus Eq. (15) becomes

$$\begin{aligned} P\big(\{\chi\}\big) = \int d\delta k P(x_0 + k_c + \delta k)(2\pi\Delta t)^{-N/2} \\ \times \exp\Bigg(-\frac{1}{2\Delta t}\sum_{n=1}^N \big[(x_n - x_{n-1}) + 2(-1)^n k_c\big]^2\Bigg) \exp\Big(-\frac{2N}{\Delta t}\delta k^2\Big) \end{aligned} \tag{17}$$

As $N \to \infty$, by the central limit theorem, k_c converges in probability to 0. On the other hand, the integrand over δk reduces to a delta function, just as we suggested. Thus the sequence of $\{x\}$ is uniquely defined, and the predictive information for $\{\chi\}$ is given by the same equations as for $\{x\}$, up to a constant, Eq. (12).

5 Solving the Model

To calculate predictive information in the model, we first calculate the Green's functions (the marginal and the conditional distributions) of Eq. (5). For this, we write the Fokker-Planck equation corresponding to the Langevin dynamics

$$\partial_t p(x,t) = \partial_x\big[x\big(x^2+\tau\big)p(x,t) + \sigma^2\partial_x\big(|x|^\alpha p(x,t)\big)\big]. \tag{18}$$

This equation immediately confirms our earlier statement that $p(x,t)=\delta(x)$ is a stationary state, stability of which depends on the strength of the noise, which in turn is controlled by α. As a result, the equation can develop a singularity near $x=0$. For any positive α, it is easy to see that the probability current at $x=0$ is zero. Physically, this corresponds to the fact that $x=0$ is an absorbing state. That is, once the system is at $x=0$, it is trapped there forever. Thus for $x_0>0$, we can consider $x(t)>0$ for any t. Further, we seek the solution for $\tau>0$, hoping further to analytically continue to the entire real axis of τ. With these caveats, we make the following simplifying transformations:

$$\bar\tau \equiv \frac{\beta^2}{\sigma^2}\hat\tau = \beta\tau/\sigma^2, \tag{19}$$

$$\hat t = t\tau/\beta, \tag{20}$$

$$\hat y \equiv y\hat\tau^{1/2} = x^{-1/\beta}\hat\tau^{1/2}, \tag{21}$$

$$f = y^{-\beta\alpha}p\big(x(y),t\big), \tag{22}$$

$$\beta = 2/(\alpha-2), \tag{23}$$

$$n = 2(\alpha-1)/(\alpha-2). \tag{24}$$

Then Eq. (18) becomes

$$\hat y^{n-1}\partial_{\hat t} f = -\partial_{\hat y}\left[\left(\hat y^n + \frac{\beta\bar\tau^{(n-3)}}{\sigma^2}\hat y^{4-n}\right)f\right] + \partial_{\hat y}\big(\hat y^{n-1}\partial_{\hat y} f\big). \tag{25}$$

The initial condition should obey $p(\hat y=0,t)=p(\hat y\to\infty,t)=0$. The former condition is a result of the inverse relationship between x and $\hat y$, while the latter is due to $x=0$ being the absorbing state.

It is important to discuss the allowed values of α at this point. From Eq. (24), n becomes divergent at $\alpha=2$. This corresponds to a large noise, which hides the phase transition. On the other hand, for large α, the noise is negligible, and the system is in an effectively deterministic regime. This happens at $n\le 3$, where the second term in Eq. (25) is suppressed as $\bar\tau\to 0$. Thus we are interested in $3<n<\infty$, which corresponds to $2<\alpha<4$. In this regime, the $\bar\tau$ term in Eq. (25) is negligibly small, and can be dropped.

With this, we notice that Eq. (25) is the radial part of the diffusion equation in n dimensions. Thus our strategy is to solve it first for n integer, hoping to analytically continue to all n later on. Assuming an integer n, we rewrite Eq. (25):

$$\partial_{\hat t} f = -nf - \hat y\partial_{\hat y} f + \frac{1}{\hat y^{n-1}}\partial_{\hat y}\big(\hat y^{n-1}\partial_{\hat y} f\big). \tag{26}$$

Therefore, $f(\hat y)$ is the radially symmetric part of the solution of the following equation

$$\partial_{\hat t} f = -nf - \hat{\mathbf y}\cdot\nabla f + \nabla^2 f. \tag{27}$$

We solve this equation in Appendix A, resulting in:

$$G(t,y,z)=C(n)z^{n-1}\left(\frac{\hat{\tau}}{2\pi(e^{2\hat{\tau}t}-1)}\right)^{n/2}$$
$$\times\int_{-1}^{1}d\lambda\exp\left(-\frac{\hat{\tau}}{2(e^{2\hat{\tau}t}-1)}\left(y^2-2yze^{\hat{\tau}t}\lambda+z^2e^{2\hat{\tau}t}\right)\right)K(\lambda),\tag{28}$$

where $K(x)$ is a kernel, which, for integer n, is the Jacobian of the n-dimensional change of variables from Cartesian to spherical coordinates. We still need to determine it for non-integer dimensions. For this, we substitute the expression of Eq. (28) in Eq. (26) (for general n) and find that it satisfies iff given by

$$\partial_\lambda^2\left[\left(1-\lambda^2\right)K(\lambda)\right]+(n-1)\partial_\lambda\left(\lambda K(\lambda)\right)=0.\tag{29}$$

To guarantee regularity at $\lambda=\pm1$ (and in analogy with the integer dimensional cases), we additionally impose the condition that $K(\pm1)=0$, leading to the solution

$$K(\lambda)=\left(1-\lambda^2\right)^{\frac{n-3}{2}}.\tag{30}$$

The normalization constant $C(n)$ can be determined from the requirement that the integral over y for a fixed z is unity when $t\to0$. In the case of an integer n, $C(n)$ is the area of the unit sphere in $n-1$ dimensions. To verify this for any value n, we need to perform the integration explicitly. To this end, it is convenient to introduce $\Delta=[(e^{2\hat{\tau}t}-1)/\hat{\tau}]^{1/2}$, and $z'=ze^{\hat{\tau}t}$. Then integrating Eq. (28), we get

$$\int_0^\infty G(t,y,z)\,dy=C(n)z^{n-1}\left(\frac{1}{\sqrt{2\pi}\,\Delta}\right)\int_0^\infty dy\exp\left(-\frac{(y-z')^2}{2\Delta^2}\right)$$
$$\times\int_{-1}^{1}(\sqrt{2\pi})^{1-n}\Delta^{1-n}\exp\left(-\frac{yz'(1-\lambda)}{\Delta^2}\right)K(\lambda)\,d\lambda.\tag{31}$$

We concentrate on the inner integral first. We perform the substitution $\xi=yz'(1-\lambda)/\Delta^2$ which leads to

$$\int_0^{2yz'/\Delta^2}\left(yz'\right)^{-\frac{n-1}{2}}(\sqrt{2\pi})^{1-n}e^{-\xi}\left[\xi\left(2-\frac{\Delta^2\xi}{yz'}\right)\right]^{\frac{n-3}{2}}d\xi$$
$$\xrightarrow[\Delta\to\infty]{}\frac{(yz)^{-\frac{n-1}{2}}}{2\pi^{(n-1)/2}}\int_0^\infty e^{-\xi}\xi^{\frac{n-3}{2}}\,d\xi=\frac{1}{2\pi^{(n-1)/2}}(yz)^{-\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right).\tag{32}$$

By dominated convergence, the limit is valid for any y and all $n>1$. (The cases $3\geq n>1$ follow from the fact that $\xi(2-\Delta^2\xi/yz')\geq\xi$ for $0<\xi\leq yz'/\Delta^2$, while the portion of the integral in Eq. 32 between $yz'/\Delta^2<\xi\leq 2yz'/\Delta^2$ converges to 0 as $\Delta\to0$). Furthermore, since yz'/Δ^2 controls the convergence in a monotonic fashion, the limit is uniform on any semi-infinite interval not containing 0. Since the convergence is dominated by a multiple of $(yz)^{-(n-1)/2}$, particularly for the values of y close to zero, we recognize the outer integral in Eq. (31) as a delta function. Therefore, in order to bring the value of Eq. (31) to unity, we need that

$$C(n)=\frac{2\pi^{(n-1)/2}}{\Gamma((n-1)/2)},\tag{33}$$

which is the area of the $n-1$ dimensional unit sphere when n is integer.

By reverting back to the original coordinate x, we can rewrite Eq. (28) and obtain the solution in these coordinates. However, for the purposes of the next section, it is more convenient to stay in the y space instead. Notice that if we make the substitutions $\tilde{p} = y^{-\alpha\beta/2} p$ in Eq. (18), we obtain

$$\partial_t \tilde{p} = -\frac{1}{\beta}\partial_y\left(\left(\hat{\tau} y + \frac{\alpha\sigma^2}{2} y^{-1} + y^{5-2n}\right)\tilde{p}\right) + \frac{\sigma^2}{\beta^2}\partial_y^2 \tilde{p}. \tag{34}$$

The advantage of $\tilde{p}$ over f calculated earlier is that $\tilde{p}$ is a probability distribution. We can immediately write its Green's function from Eq. (28) since $\tilde{p}(t, y) = y^{n-1} f(t, y)$:

$$\begin{aligned}\tilde{G}(t, y, z) = {} & C(n)(y)^{n-1}\left(\frac{\hat{\tau}}{2\pi(e^{2\hat{\tau}t} - 1)}\right)^{n/2} \\ & \times \int_{-1}^{1} d\lambda \exp\left(-\frac{\hat{\tau}}{2(e^{2\hat{\tau}t} - 1)}\left(y^2 - 2yze^{\hat{\tau}t}\lambda + z^2 e^{2\hat{\tau}t}\right)\right) K(\lambda).\end{aligned} \tag{35}$$

This is the main result of this section, which we will use in order calculate predictive information for our model. One can verify by explicit substitution that the expression in Eq. (35) satisfies the Fokker-Planck equation, Eq. (25), and it reduces to a delta function as $t \to 0$. Thus it represents the conditional distribution of y given z.

6 Predictive Information for the Model

Predictive information is reparameterization invariant. Thus we can calculate it for y instead of x and use the expression, Eq. (35), when applying the Eq. (12) to our model. Without loss of generality, we assume that the initial condition is a delta function. Then the continuous form of Eq. (12) is

$$I_{\text{pred}}(t) = \left\langle \log_2 \frac{\tilde{G}(\tilde{t}, y, z)}{\tilde{G}(t + \tilde{t}, y, w)} \right\rangle, \tag{36}$$

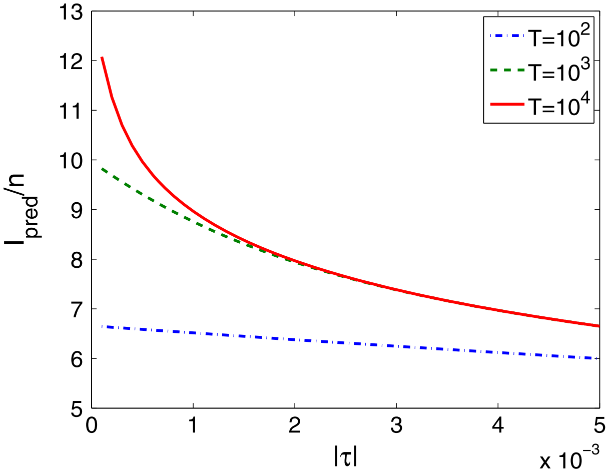
where w, z, and y are the values of the observable at times 0, $t = (N-1)\Delta t$, and $T \equiv t + \tilde{t} = (N+L)\Delta t$ respectively, i. e., $w = x_0^{-1/\beta}$, $z = x_{N-1}^{-1/\beta}$, and $y = x_{N+L}^{-1/\beta}$. Equation (36) involves an integral with complex time and $\hat{\tau}$- dependences. In the following, we would like to find the leading orders of these dependences. Defining $\Delta(t) = [(e^{2\hat{\tau}t} - 1)/\hat{\tau}]^{1/2}$ (cf. Eq. (28)), it is also convenient to introduce $\Xi(t; \lambda, y, w) = \exp[(y^2 - 2ywe^{\hat{\tau}t}\lambda + w^2 e^{2\hat{\tau}t})/(2\Delta(t)^2)]$, so that Eq. (35) takes on the form

$$\tilde{G}(t, y, z) = C(n)\frac{(2\pi)^{-n/2}}{\Delta(t)^n}\int_{-1}^{1} d\lambda\, K(x)\, \Xi(t; \lambda, y, z). \tag{37}$$

Then Eq. (36) becomes

$$\begin{aligned} I_{\text{pred}}(t) = {} & n\log_2 \frac{\Delta(T)}{\Delta(\tilde{t})} + \left\langle \log_2 \int_{-1}^{1} d\lambda K(\lambda)\Xi(\tilde{t}; \lambda, y, z)\right\rangle \\ & - \left\langle \log_2 \int_{-1}^{1} d\lambda K(\lambda)\Xi(T; \lambda, z, w)\right\rangle. \end{aligned} \tag{38}$$

Fig. 1 A plot of I_{pred}/n for different values of $\tau < 0$ at different times T for a fixed $\tilde{t} = 1$



In Appendix C, we show that the last two terms in Eq. (38) are asymptotically constant when $T \to \infty$ if t is large and $\hat{\tau}$ is small. Therefore, to the leading order, predictive information is

$$I_{\text{pred}}(t) \approx n \log_2 \frac{\Delta(T)}{\Delta(\tilde{t})} = n \log_2 \frac{\exp[2\hat{\tau}(t+\tilde{t})] - 1}{\exp(2\hat{\tau}\tilde{t}) - 1}. \tag{39}$$

At the critical point, when the absorbing state is just starting to emerge, $\hat{\tau} \to 0$, this expression reduces to

$$I_{\text{pred}}(t) \approx n \log_2 \frac{t+\tilde{t}}{\tilde{t}}. \tag{40}$$

This logarithmic growth with the system size t has been anticipated for a critical point in Ref. [15], but has not been calculated before for any nonequilibrium stochastic dynamical system. A plot of Eq. (39) is given for different parameter values in Fig. 1.

Notice that the prefactor $n = 2(\alpha - 1)/(\alpha - 2)$ increases with the effect of the noise, which corresponds to more of partially predictable variability in the dynamics, and hence to an intuitively higher complexity. Further, as $\alpha \to 2$, or $n \to \infty$, the critical point is smeared and a non-absorbing steady state emerges for all values of τ. In this limit, the subleading term in the entropy diverges (or, more precisely, becomes extensive), and hence it would cancel out in the difference of entropies in Eq. (2), leading to $I_{\text{pred}}(t) = \text{const}$. We can see this from Eq. (39), since, for large n and negative τ, we have $\hat{\tau} \sim -n$ and $I_{\text{pred}}(t) \approx n \log_2[1 + \exp(-n\tilde{t})] \to 0$ as $n \to \infty$ (notice the importance of the order of the two limits, $n \to \infty$ and $\hat{\tau} \to 0$). Equation (39) also allows calculation of the asymptotic away from the phase transition. For large negative $\hat{\tau}$, $I_{\text{pred}}(t) = \text{const}$. For large positive τ, $I_{\text{pred}}(t) \propto t$, since perfect prediction is possible in the absorbing state. This leads to $\lim_{t\to\infty} I_{\text{pred}}(t)/t \neq 0$. This is in contrast to the equilibrium analysis of [15, 16], where I_{pred} is always sublinear. This is also a direct result of the continuous nature of our problem: consecutive measurements allow higher accuracy description of a known real-valued state of the system (specifically, $x = 0$), instead of predicting its (unknown) state. This divergence is not related to the UV divergence, $\sim \log(\Delta t)$, discussed earlier. We expect it to be present in many model with absorbing states, aging, or other nonequilibrium phenomena.

Taken together, these results illustrate that the logarithmic divergence of predictive information correctly captures the existence of the phase transition (emergence of the absorbing state) at $\tau \to 0$.

7 Discussion

Predictive information was introduced in Ref. [15] as information between the past and the future of a time series, or between left and right parts of a physical system. It was argued, in particular, that the behavior of predictive information as the system size grows can signal existence of a phase transition. As an example, Ref. [15] calculated the information numerically for an equilibrium long-range one-dimensional Ising magnet. In the current work, we argue that predictive information can be used as a general order parameter in more complicated scenarios, such as in nonequilibrium contexts, where traditional symmetry arguments fail to identify low-order correlation functions that can serve this role. For the first time, we calculate predictive information for a nonequilibrium Markov process, which exhibits a phase transition at certain values of parameters, where a logarithmic divergence of I_{pred} develops. Intuitively, this logarithmic divergence is a byproduct of a transition from $I_{\text{pred}} = \text{const}$ to $I_{\text{pred}} \sim t$, just like critical phenomena in general are byproducts of transitions between phases. In equilibrium systems, the constant and the linearly diverging phases are equivalent, since linear contributions cancel out in the definition of I_{pred}, Eq. (4). This is not true anymore in our nonequilibrium problem with aging, and yet the logarithmic divergence of predictive information still correctly captures this phase transition.

While the logarithmic behavior of I_{pred} at criticality has been observed previously in Ref. [16], and stronger divergences were speculated to exist for certain transitions in glassy systems, it has been unclear how the predictive information transitions from the divergence *at* a critical point to a constant *far away* from it. Our calculations reveal the exact form of this smooth crossover. To our knowledge, this has not been calculated before, either for equilibrium or for nonequilibrium systems.

Mathematically, the logarithmic divergence can be explained since the propagator, Eq. (35), can be written as $\tilde{G}(t, y, z) \approx \tilde{G}(y/t, z/t)$ when $\hat{\tau} = 0$. In fact, we conjecture that predictive information will exhibit an asymptotically logarithmic behavior for any model possessing power-law time dependencies, with exponents drawn from a bounded set, even if such dependencies are not limited to the two-point correlators. This parallels the relation between the power-law scaling and the collapse of the spectrum of the dynamical operator. Detection of the latter is not always possible from the measurements of a few correlations functions, particularly when there is still a gap. However, we believe that it would still lead to a logarithmic behavior of the predictive information. Therefore, divergence of I_{pred} may answer another question. Namely, one typically associates critical phenomena with power-law scaling of the two-point correlation. As we mentioned in the Introduction, this is not applicable to a wide range of problems. Thus there is no universally accepted definition of criticality. Our results suggest that logarithmic (and, more generally, non-analytic) divergence of I_{pred} may be used as a basis for such definition. As we calculate I_{pred} in more systems, we can verify the extensivity of the latter across the spectrum of definitions of criticality.

One important technical difference between this work and the previous ones is the introduction of an additional "renormalization" scale, L or $\tilde{t}$, in the definition of predictive information, so that the information is calculated between the past and the future that are separated by a finite distance. This removed the ultraviolet divergences associated with information at the interface between the past and the future of a trajectory. While this modification was precipitated by the continuous time/space nature of the stochastic process, we believe that it will solve additionally difficulties with application of predictive information ideas to systems with more than one dimension. Indeed, there the main problem is that the interface between two parts of a system diverges with the system size, and hence the interfacial contribution to predictive information diverges even away from a critical point. This will not happen if direct interfaces are eliminated.

In summary, in this paper, we provide the first example of a direct analytical calculation of predictive information for a nonequilibrium stochastic process. This example argues further for using predictive information as a general order parameter for studying phase transitions.

Acknowledgements This work has been supported in part by a James S. McDonnell Foundation Complex Systems Grant No. 220020321. We would like to thank HGE Hentschel for stimulating discussions.

Appendix A: Calculating the Green's Function

Green's function of Eq. (27) is found easier in the Cartesian coordinates, and the radial component can be extracted afterwards. Thus we look for the Green's function of the form

$$G(\hat{t}; \hat{\mathbf{y}}, \hat{\mathbf{z}}) = \prod_{i}^{n} G_1(t; \hat{y}_i, \hat{z}_i) \tag{41}$$

where $G_1(\hat{t}; \hat{y}_i, \hat{z}_i)$ is the one dimensional Green's function, satisfying

$$\partial_{\hat{t}} G_1 = -G_1 - \hat{y}\partial_{\hat{y}} G_1 + \partial_{\hat{y}}^2 G_1 + \delta(\hat{t}, \hat{y} - \hat{z}). \tag{42}$$

To solve Eq. (42), it is convenient to consider $\tilde{G}_1 = e^{\hat{t}} G_1$, where $\tilde{G}_1$ satisfies

$$\partial_{\hat{t}} \tilde{G}_1(\hat{t}; \hat{y}, \hat{z}) - \partial_{\hat{y}}^2 \tilde{G}_1(\hat{t}; \hat{y}, \hat{z}) + \hat{y}\partial_{\hat{y}} \tilde{G}_1(\hat{t}; \hat{y}, \hat{z}) = \delta(\hat{t}, \hat{y} - \hat{z}). \tag{43}$$

As usual, we transform into Fourier space:

$$i\omega \tilde{G}_1 + k^2 \tilde{G}_1 - \partial_k (k \tilde{G}_1) = e^{-ik\hat{z}}. \tag{44}$$

If we use the integral multiplier

$$\mu = \exp\bigl(-\bigl(i\omega \ln k + k^2/2\bigr)\bigr), \tag{45}$$

we obtain the following simplified form of Eq. (44)

$$-\partial_k (k \mu \tilde{G}_1) = \mu e^{-ik\hat{z}}. \tag{46}$$

Since we are looking for a smooth solution, we expect $\tilde{G} = 0$ as $k \to \infty$. Therefore, the correct solution of the above equation is in the form

$$\tilde{G}_1(\omega, k, \hat{z}) = k^{-1} \mu^{-1} \int_k^{\infty} e^{-ik'\hat{z}} e^{-(iw \ln k' + k'^2/2)}\, dk'. \tag{47}$$

Inverting back to the time coordinate, we obtain

$$\tilde{G}_1(\hat{t}, k, \hat{z}) = e^{k^2/2} k^{-1} \int_k^{\infty} e^{-ik'\hat{z}} e^{-k'^2/2} \delta\big(\hat{t} - \ln k' + \ln k\big)\, dk'. \tag{48}$$

Now performing the delta function integration, we are left with

$$\tilde{G}_1(\hat{t}, k, \hat{z}) = e^{k^2/2} e^{\hat{t}} e^{-ike^{\hat{t}}\hat{z} - k^2 e^{2\hat{t}}/2}. \tag{49}$$

This is simply a Gaussian function, and the transformation back to the $\hat{y}$ coordinate leaves us with

$$G_1(\hat{t}, \hat{y}, \hat{z}) = e^{-\hat{t}} \tilde{G}_1(\hat{t}, \hat{y}, \hat{z}) = \big[2\pi\big(e^{2\hat{t}} - 1\big)\big]^{-1/2} \exp\bigg(-\frac{1}{2}\frac{(\hat{y} - e^{\hat{t}}\hat{z})^2}{e^{2\hat{t}} - 1}\bigg). \tag{50}$$

We would like to extract the full dependence of the above solution on $\hat{\tau}$. For normalization purposes, it is also convenient to multiply by $\hat{\tau}^{1/2}$. Thus rescaling back to the t and y coordinates results in

$$G_1(t, y, z) = \bigg(\frac{\hat{\tau}}{2\pi(e^{2\hat{\tau}t} - 1)}\bigg)^{1/2} \exp\bigg(-\frac{\hat{\tau}}{2}\frac{(y - e^{\hat{\tau}t}z)^2}{e^{2\hat{\tau}t} - 1}\bigg). \tag{51}$$

This finally results in an expression for the Green's function of Eq. (42), which in turn gives the Green's function of Eq. (27) in Cartesian coordinates. Now, to obtain the solution of Eq. (26), we need to revert back to spherical coordinates. The resulting expression when n is integer suggests that we look for $G(t, y, z)$ in the following form

$$\begin{aligned} G(t, y, z) = C(n) z^{n-1} &\bigg(\frac{\hat{\tau}}{2\pi(e^{2\hat{\tau}t} - 1)}\bigg)^{n/2} \\ &\times \int_{-1}^{1} d\lambda \exp\bigg(-\frac{\hat{\tau}}{2(e^{2\hat{\tau}t} - 1)}\big(y^2 - 2yze^{\hat{\tau}t}\lambda + z^2 e^{2\hat{\tau}t}\big)\bigg) K(\lambda). \end{aligned} \tag{52}$$

Here $K(x)$ is a kernel, which, for integer n, is the Jacobian of the n-dimensional change of variables from Cartesian to spherical coordinates. It is still undetermined for non-integer dimensions.

Appendix B: Identification of Terms Dominating Convergence

In the main text, we argued that it is justifiable to drop the y^{4-n} term in Eq. (25), or equivalently, the y^{5-2n} term in Eq. (34). In essence, the bulk of the solution is supported away from $y = 0$, while this term is quickly suppressed for $n > 3$. Without this (generally) non-integer power, we were able to calculate exactly predictive information for our model. Whatever the contributions the full solution might add, they are of lower order than the leading term in Eq. (39). Nonetheless, this term is crucial since it keeps the full solution physical by guaranteeing its convergence faster than any power as $y \to 0$ ($x \to \infty$ in the x space). In this appendix, we will make the arguments a bit more precise.

Our approach is of the maximum principle type, which is employed abundantly in the theory of partial differential equations. We present the arguments in a general setting, not limited to the confines of our model. Our focus is on equations of the type

$$\partial_t F(t, y) = -g(y)\partial_y F(t, y) + \partial_y^2 F(t, y), \qquad y > 0. \tag{53}$$

F is the cumulative probability $\int_0^y f(t,y')\,dy'$ of a distribution f satisfying a Fokker-Planck equation with constant noise and a force $g(y)$. We will assume that around $y \sim 0$, g is positive and behaves as $1/y^\alpha$ with $\alpha > 1$. We start by providing a sort of a zero value "eigenvector", i.e. a solution of the equation

$$0 = -g(y)\partial_y F_0(y) + \partial_y^2 F_0(y). \tag{54}$$

It is straightforward to see that Eq. (54) is solved by

$$F_0(y) = \int_0^y dy' \exp\biggl(\int_{y_0}^{y'} dy'' g\bigl(y''\bigr)\biggr), \tag{55}$$

where y_0 is any positive value. It follows that $F_0(y) \sim \exp(-1/y^{\alpha-1})$, thus it converges to zero, together with all of its derivatives.

The solution, Eq. (54), is non-normalizable, and it is, therefore, not a true eigenvector. However, we can use it to bound normalizable solutions of Eq. (53). That is, we will show that if initial conditions are bounded everywhere by a multiple of F_0 (e. g., if their support does not include 0), then the solution $F(t,y)$ remains bounded for all times, and it will, therefore, have all derivatives zero at $y = 0$. This implies that the exact solution of Eq. (18) indeed has a finite tail, and this is all due to the third term in Eq. (34). By imposing the requirement that this term diverges faster than $1/y$, we obtain $n > 3$, or equivalently $\alpha < 4$.

In order to demonstrate that $F(t,y) \le F_0(y)$ if $F(0,y) \le F_0(y)$, we will first show the following.

If $\tilde F(t,y)$ satisfies the boundary conditions $\tilde F(t,0) = 0$ and $\tilde F(t,L) \ge 1/2$, for some $L > 0$, together with the initial condition $\tilde F(0,y) \ge 0$, then $\tilde F$ remains non-negative for all times if it also satisfies the following equation:

$$\partial_t \tilde F(t,y) = -\gamma \tilde F(t,y) - g(y)\partial_y \tilde F(t,y) + \partial_y^2 \tilde F(t,y), \quad \gamma > 0. \tag{56}$$

Proof Assume a negative minimum of $\tilde F(t_0,y_0) < -\epsilon$ at some time t_0 and point y_0. Clearly, $0 < y_0 < L$. Then, at y_0:

$$\begin{aligned}\partial_t \tilde F(t_0,x_0) &= \partial_y^2 \tilde F(t_0,x_0) - g(y_0)\partial_y \tilde F(t_0,y_0)\\ &\quad - \gamma \tilde F(t_0,y_0) \ge \partial_y^2 \tilde F(t_0,x_0) + \gamma\epsilon > 0.\end{aligned} \tag{57}$$

This implies that there is a $\delta > 0$ such that $\tilde F < -\epsilon$ at some points y, for all $t_0 - \delta < t < t_0$. Let $\tilde t$ be the infimum of the set of all times for which $\tilde F < -\epsilon$ at some point. Take a sequence $\{t_n\}$ which converges to $\tilde t$ and a sequence $\{y_n\}$ such that $\tilde F(t_n,y_n) < -\epsilon$. Since $0 < \{y_n\} < L$, we can assume that it converges to some $\tilde y \ne 0$. Thus, $\tilde F(\tilde t,\tilde y) < -\epsilon$. By applying Eq. (57) again, we obtain that this is possible only if $\tilde t = 0$, which, in turn, is impossible because of the initial conditions. □

Notice that the positivity of $\tilde F$ immediately implies the positivity of F since there is a one-to-one mapping between the solutions of Eqs. (53) and (56) given by $\tilde F \exp(\gamma t) = F$. If we apply this to $\Delta F(t,y) \equiv F_0(t,y) - F(t,y)$, then $F_0(t,y) \ge F(t,y)$ for all times t, as long as this is true for $t = 0$, just as we claimed earlier. We end with a comment regarding the boundary condition requirement at $y = L$. If F_0 is non-normalizable, then this condition is trivially satisfied. Otherwise, this condition is a byproduct of the uniqueness requirements of the solution. Therefore, the approximate solution, Eq. (28), is an upper bound on the exact solution of Eq. (18).

Appendix C: Bounding Subleading Terms in Predictive Information

While we have not been able to obtain a closed form expression for all terms in Eq. (38), we can nonetheless provide asymptotically finite bounds on them. We will rely on the basic structure of the solution, Eq. (37), and repeated applications of the Jensen's inequality.

Starting with the full expression in Eq. (38), we would like to start by providing the following bounds for $z > 0$ and $\theta > 1$, $\vartheta > 0$:

$$A(\theta,\vartheta) + B(\theta,\vartheta) z^{\theta} \geq \int_0^{\infty} y^{\theta} (y-z)^{\vartheta} e^{-(y-z)^2/2} \geq a(\theta,\vartheta) z^{\theta-1} + b(\theta,\vartheta) z^{\theta}. \tag{58}$$

Here $A,\ B,\ a,\ b$ are positive functions of θ and ϑ only. It is useful to normalize the kernel $K(\lambda)$. Thus we define

$$\kappa = \int_{-1}^{1} K(\lambda)\, d\lambda = 2^{n-2} \frac{\Gamma(\frac{n-1}{2})^2}{\Gamma(n-1)}, \tag{59}$$

where the last equality contains the usual Gamma function. We now can provide an upper bound on the integral terms in Eq. 38. By using the fact that $x \log(x)$ is a convex function, we obtain

$$\begin{aligned}
&C^{-1}(2\pi)^{n/2} \left\langle \log_2 \int_{-1}^{1} d\lambda K(\lambda) \varXi(T;\lambda,y,w) \right\rangle - C^{-1}(2\pi)^{n/2} \log_2 \kappa \\
&\quad = \kappa \int_0^{\infty} dy \frac{y^{n-1}}{\Delta^n(T)} \int_{-1}^{1} d\lambda \frac{K(\lambda)}{\kappa} \varXi(T;\lambda,y,w) \log_2 \int_{-1}^{1} d\lambda' \frac{K(\lambda')}{\kappa} \varXi\big(T;\lambda',y,w\big) \\
&\quad \leq \kappa \int_0^{\infty} dy \frac{y^{n-1}}{\Delta^n(T)} d\lambda \frac{K(\lambda)}{\kappa} \varXi(T;\lambda,y,w) \log_2 \varXi(T;\lambda,y,w) \\
&\quad \leq -(1/2)\log_2(e) \int_{-1}^{1} d\lambda K(\lambda) \Bigg[a(n-1,2)\lambda^{n-2} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n-2} \\
&\qquad + b(n-1,2)\lambda^{n-1} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n-1} + a(n-1,0)\big(1-\lambda^2\big)\lambda^{n-2} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n} \\
&\qquad + b(n-1,0)\big(1-\lambda^2\big)\lambda^{n-1} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n+1} \Bigg] e^{-\frac{1-\lambda^2}{2\Delta(T)^2} w^2 e^{2\hat{\tau}T}}.
\end{aligned} \tag{60}$$

Similarly, utilizing the concavity of $\log(x)$, we can write a lower bound on the expectation value

$$\begin{aligned}
&C^{-1}(2\pi)^{n/2} \left\langle \log_2 \int_{-1}^{1} d\lambda K(\lambda) \varXi(T;\lambda,y,w) \right\rangle - C^{-1}(2\pi)^{n/2} \log_2 \kappa \\
&\quad \geq -(1/2)\log_2(e) \int_{-1}^{1} d\lambda K(\lambda) \Bigg[A(n+1,0) + B(n+1,0)\lambda^{n+1} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n+1} \\
&\qquad \times A(n-1,0) \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{2} + B(n-1,0)\lambda^{n-1} \left(\frac{w e^{\hat{\tau}T}}{\Delta(T)}\right)^{n+1} \Bigg] e^{-\frac{1-\lambda^2}{2\Delta^2(T)} w^2 e^{2\hat{\tau}T}}.
\end{aligned} \tag{61}$$

Therefore, we have obtained bounds on the third term in Eq. (38) that are polynomial in $e^{\hat{\tau}T}/\Delta(T)$. The latter is, in turn, a bounded function of $T = t + \tilde{t}$. Indeed, it is straightforward to show that $e^{\hat{\tau}T}/\Delta(T) \leq \sqrt{|\hat{\tau}|} + \sqrt{1/T}$. Therefore, these bounds are asymptotically constant (as $T \to \infty$) and either $\mathcal{O}(1)$ or $\mathcal{O}(\sqrt{|\hat{\tau}|})$. We can use these bounds on the second term of Eq. (38) by simply replacing w by z and T by $\tilde{t}$ in Eqs. (60) and (61). The resulting expressions need to be averaged over z, which requires estimating quantities of the form

$$\begin{aligned} L \le \int_0^\infty dz \frac{z^{n-1}}{\Delta^n(t)} z^m \left(\frac{e^{\hat{\tau}\tilde{t}}}{\Delta(\tilde{t})}\right)^m e^{-\frac{1-\lambda^2}{2\Delta^2(\tilde{t})} e^{2\hat{\tau}\tilde{t}} z^2} \\ \times \int_{-1}^{1} d\tilde{\lambda} K(\tilde{\lambda}) \exp\left(-\frac{1}{2\Delta(t)^2}\left(z^2 - 2zw\tilde{\lambda}e^{\hat{\tau}t} + w^2 e^{2\hat{\tau}t}\right)\right) \le U, \end{aligned} \tag{62}$$

where m is a positive number. By using Eq. (58) again, we can obtain an upper and a lower bound on this expression. It is convenient to introduce $\eta^2 = (1-\lambda^2)e^{2\hat{\tau}\tilde{t}}\Delta^2(t)/\Delta^2(\tilde{t})$. Then, after some algebra, we obtain the following two bounds: an upper bound

$$\begin{aligned} U = \int_{-1}^{1} d\tilde{\lambda} K(\tilde{\lambda}) \left[\frac{\eta^2}{1-\lambda^2}\right]^{m/2} \left(1+\eta^2\right)^{-(n+m)/2} \Bigg[A(n+m-1,0) \\ + B(n+m-1,0)\left(\frac{\tilde{\lambda} w e^{\hat{\tau}t}}{\Delta(t)(1+\eta^2)^{1/2}}\right)^{n+m-1}\Bigg] \exp\left(-\frac{1}{2}\left(1-\frac{\tilde{\lambda}^2}{1+\eta^2}\right)\frac{w^2 e^{2\hat{\tau}t}}{\Delta^2(t)}\right), \end{aligned} \tag{63}$$

and a lower bound

$$\begin{aligned} L = \int_{-1}^{1} d\tilde{\lambda} K(\tilde{\lambda}) \left[\frac{\eta^2}{1-\lambda^2}\right]^{m/2} \left(1+\eta^2\right)^{-(n+m)/2} \left(\frac{\tilde{\lambda} w e^{\hat{\tau}t}}{\Delta(t)(1+\eta^2)^{1/2}}\right)^{n+m-2} \\ \times \left[a(n+m-1,0) + b(n+m-1,0)\left(\frac{\tilde{\lambda} w e^{\hat{\tau}t}}{\Delta(t)(1+\eta^2)^{1/2}}\right)\right] \\ \times \exp\left(-\frac{1}{2}\left(1-\frac{\tilde{\lambda}^2}{1+\eta^2}\right)\frac{w^2 e^{2\hat{\tau}t}}{\Delta^2(t)}\right). \end{aligned} \tag{64}$$

Notice that, for $\hat{\tau} \ge 0$, $\eta \to \infty$ as $t \to \infty$, while both bounds in Eqs. (63) and (64) are of order $\mathcal{O}(\eta^{-m/2})$, therefore they are asymptotically constant. For $\hat{\tau} < 0$, Eqs. (63) and (64) are controlled by $\mathcal{O}(|\hat{\tau}|^{m/2})$. This implies that the second term in Eq. (38) is also bounded around the critical point, independently of $\hat{\tau}$. This completes the proof that the terms we dropped in Eq. (38) do not contribute to the leading order of predictive information.

References

1. Brazhkin, V., Trachenko, K.: What separates a liquid from a gas? Phys. Today **65**(11), 68–69 (2012)
2. Tchernookov, M., Warmflash, A., Dinner, A.: Field theoretic treatment of an effective action for a model of catalyzed autoamplification. Phys. Rev. E **81**, 011112 (2010)
3. Van Wijland, F., Oerding, K., Hilhorst, H.J.: Wilson renormalization of a reaction–diffusion process. Physica A **251**, 179 (1998)
4. Garrahan, J., Jack, R., Lecomte, V., Pitard, E., van Duijvendijk, K., van Wijland, F.: Dynamical first-order phase transition in kinetically constrained models of glasses. Phys. Rev. Lett. **98**, 195702 (2007)
5. Biroli, G., Bouchaud, J.: Diverging length scale and upper critical dimension in the Mode-Coupling Theory of the glass transition. Europhys. Lett. **67**, 21 (2007)
6. Binder, K., Landau, D.P.: Critical phenomena at surfaces. Physica A **163**(1), 17 (1990)
7. Le Doussal, P., Wiese, K., Chauve, P.: Functional renormalization group and the field theory of disordered elastic systems. Phys. Rev. E **69**, 026112 (2004)

8. Lecomte, V., Appert-Rolland, C., Van Wijland, F.: Thermodynamic formalism for systems with Markov dynamics. J. Stat. Phys. **127**, 51 (2007)
9. Kurchan, J., Levine, D.: Order in glassy systems. J. Phys. A **44**(3), 035001 (2010)
10. Garrahan, J.P., Jack, R.L., Lecomte, V., Pitard, E., van Duijvendijk, K., van Wijland, F.: First-order dynamical phase transition in models of glasses: an approach based on ensembles of histories. J. Phys. A **42**(7), 075007 (2009)
11. Hedges, L.O., Jack, R.L., Garrahan, J.P., Chandler, D.: Dynamic order-disorder in atomistic models of structural glass formers. Science **323**(5919), 1309–1313 (2009)
12. Gaveau, B., Schulman, L.: Theory of nonequilibrium first-order phase transitions for stochastic dynamics. J. Math. Phys. **39**, 1517 (1998)
13. Castelnovo, C., Chamon, C., Sherrington, D.: Quantum mechanical and information theoretic view on classical glass transitions. Phys. Rev. B **81**(18), 184303 (2010)
14. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. Univ Illinois Press, Urbana (1998)
15. Bialek, W., Nemenman, I., Tishby, N.: Predictability, complexity, and learning. Neural Comput. **13**, 2409–2463 (2001)
16. Bialek, W., Nemenman, I., Tishby, N.: Complexity through nonextensivity. Physica A **302**, 89–99 (2001)
17. Grassberger, P.: Toward a quantitative theory of self-generated complexity. Int. J. Theor. Phys. **25**(9), 907–938 (1986)
18. Gaspard, P.: Chaos, Scattering and Statistical Mechanics, vol. 9. Cambridge University Press, Cambridge (2005)
19. Shaw, R.: The Dripping Faucet as a Model Chaotic System. Aerial Press, Santa Cruz (1984)
20. Crutchfield, J., Feldman, D.: 2001, Regularities unseen, randomness observed: levels of entropy convergence. Preprint, cond-mat/0102181
21. Shalizi, C., Crutchfield, J.: Computational mechanics: pattern and prediction, structure and simplicity. J. Stat. Phys. **104**(3), 817–879 (2001)
22. Vereshchagin, N., Vitányi, P.: Kolmogorov's structure functions and model selection. IEEE Trans. Inf. Theory **50**(12), 3265–3290 (2004)
23. Vitányi, P.: Meaningful information. IEEE Trans. Inf. Theory **52**(10), 4617–4626 (2006)
24. Calabrese, P., Cardy, J.: Entanglement entropy and conformal field theory. J. Phys. A **42**(50), 504005 (2009)
25. Still, S., Sivak, D.A., Bell, A.J., Crooks, G.E.: Thermodynamics of prediction. Phys. Rev. Lett. **109**(12), 120604 (2012)
26. Spohn, H.: Long range correlations for stochastic lattice gases in a non-equilibrium steady state. J. Phys. A **16**(18), 4275 (1983)
27. Tailleur, J., Kurchan, J., Lecomte, V.: Mapping out-of-equilibrium into equilibrium in one-dimensional transport models. J. Phys. A **41**(50), 505001 (2008)
28. Elderfield, D., Vvedensky, D.: Non-equilibrium scaling in the Schlogl model. J. Phys. A, Math. Gen. **18**(13), 2591 (1999)
29. Monthus, C.: Non-equilibrium steady states: maximization of the Shannon entropy associated with the distribution of dynamical trajectories in the presence of constraints. J. Stat. Mech. **2011**(03), P03008 (2011)
30. Gaspard, P.: Time-reversed dynamical entropy and irreversibility in Markovian random processes. J. Stat. Phys. **117**(3), 599 (2004)
31. Jones, G.: On the Markov chain central limit theorem. Probab. Surv. **1**, 299 (2004)