Entropy **2014**, *16*, 4713-4748; doi:10.3390/e16094713

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Information Anatomy of Stochastic Equilibria

Sarah Marzen [1,*] and James P. Crutchfield [2,*]

[1] Department of Physics, University of California at Berkeley, Berkeley, CA 94720, USA

[2] Complexity Sciences Center, Department of Physics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

* Authors to whom correspondence should be addressed; E-Mails: smarzen@berkeley.edu (S.M.); chaos@ucdavis.edu (J.P.C.)

Received: 17 March 2014; in revised form: 3 August 2014 / Accepted: 19 August 2014 / Published: 25 August 2014

Abstract: A stochastic nonlinear dynamical system generates information, as measured by its entropy rate. Some—the ephemeral information—is dissipated and some—the bound information—is actively stored and so affects future behavior. We derive analytic expressions for the ephemeral and bound information in the limit of infinitesimal time discretization for two classical systems that exhibit dynamical equilibria: first-order Langevin equations (i) where the drift is the gradient of an analytic potential function and the diffusion matrix is invertible and (ii) with a linear drift term (Ornstein–Uhlenbeck), but a noninvertible diffusion matrix. In both cases, the bound information is sensitive to the drift and diffusion, while the ephemeral information is sensitive only to the diffusion matrix and not to the drift. Notably, this information anatomy changes discontinuously as any of the diffusion coefficients vanishes, indicating that it is very sensitive to the noise structure. We then calculate the information anatomy of the stochastic cusp catastrophe and of particles diffusing in a heat bath in the overdamped limit, both examples of stochastic gradient descent on a potential landscape. Finally, we use our methods to calculate and compare approximations for the time-local predictive information for adaptive agents.

Keywords: Langevin equation; entropy rate; ephemeral information; bound information; time-local predictive information

---

1. Introduction

If we track the position of a particle diffusing on an unchanging potential long enough, we can estimate the probability of observing a sequence of positions [1]. From that, we can quantitatively answer questions about the process's behavior using a range of information statistics that answer specific questions:

- How random is it? The entropy rate h_μ, which is the entropy in the present observation conditioned on all past observations [2].
- What must be remembered about the past in order to optimally predict the future? The causal states, which are groupings of pasts that lead to the same probability distribution over future trajectories [3,4].
- How much memory is required to store these causal states? The statistical complexity C_μ, or the entropy of the causal states [3].
- How much of the future is predictable from the past? The excess entropy $\mathbf{E}$, which is the mutual information between the past and the future [5].
- How much of the generated information (h_μ) is relevant to predicting the future? The bound information b_μ, which is the mutual information between the present and future observations conditioned on all past observations [6].
- How much of the generated information is useless (neither affects future behavior nor contains information about the past)? The ephemeral information r_μ, which is the entropy in the present observation conditioned on all past and future observations [6].

These informational quantities usually cannot be deduced from a bifurcation diagram, so we see them as providing a complementary view of a process's structure and behavior.

In applications, such informational characterizations of a time series are useful for monitoring good sensory coding [7], cognitive modalities [8], brain coherence [9], hidden Markov model structural inference [10], action policies of autonomous agents [11,12], structure in disordered materials [13,14], dynamical phase transitions [15,16] and intrinsic information processing in deterministic chaos [17,18] and cellular automata [19,20].

Here, we focus on continuous stochastic nonlinear dynamical systems, the theory for which has a long and venerable history, has met with a number of successful predictions, and has identified a number of principles describing how noise interacts with nonlinearity [21]. For nonlinear systems transitioning to chaos, to take just one example, noise plays the role of a "disordering" field, just as the magnetic field is an ordering field for spin systems at critical transitions [22,23]. Though their history substantially predates that of the wide range of complex systems applications just cited, relatively fewer analyses of their information processing components—their information anatomy—have been carried out. As a start, we demonstrate how to calculate the quantities above for continuous-time, continuous-state stochastic nonlinear systems exhibiting dynamical equilibria, yielding intuition for the properties these measures capture in simpler, and perhaps more familiar, physical models.

Throughout, we focus on a ubiquitous and simple nonlinear generative model: stochastic gradient descent or, in other words, diffusion on a potential surface. We assume infinite precision in our observation of the state space. The first calculation assumes that the diffusion matrix is invertible and

drift is analytic; the second assumes that the drift term is linear, but allows for a noninvertible diffusion matrix. All calculations assume that the time between measurements is nonzero, but arbitrarily small, and that all derived information anatomy quantities are finite at finite temporal coarse-graining.

There are alternative ways to frame information analyses of continuous stochastic processes [24]. The one we take is rather prosaic, paralleling the "physics" approach laid out by Gaspard and Wang [25], who coarse-grain time at a finite, but small, time scale τ and state-space at a similar spatial scale ϵ. This discretizes the calculations and then one takes the limits $\tau \to 0$ and $\epsilon \to 0$. Crucially, the limits often reveal divergences in the informational quantities. For example, it is well known that the (ϵ, τ)-entropy of a broad family of continuous stochastic processes diverges [25]. However, as Gaspard and Wang demonstrate and as is familiar in other fundamental physics domains, the form of the divergences captures important structural properties. The main deviation here from their approach is that we employ the Shannon differential entropy to side-step state-space (ϵ) coarse-graining.

An alternative, and insightful, framing considers the divergences to be unnatural; in particular, with naive coarse-graining, it is difficult to establish ergodic theorems key to information theory. In this view, the main concern translates into a search for tractable definitions of information measures that finitely quantify information processing in continuous stochastic systems. To address divergences, one investigates a given stochastic process relative to Brownian motion. In a crude sense, the known Brownian base case carries the divergences. To factor them out of the given process, one employs Girsanov's theorem to transform the given process to a canonical Brownian motion with the same diffusion [26]. Properties of the transformation then characterize the given process's informational properties; for example, giving a relative entropy rate. This strikes us as an important avenue for future investigation; one that, to be clear, is not yet completed, as far as we know, and one that eventually will be related to the more prosaic, physics framing that we address here.

To get started, background is given in Section 2. Results are presented in Section 3 and stated more succinctly in Table 1. To illustrate how to apply those formulae, we calculate the information anatomy of the stochastic cusp catastrophe in Section 4.1 and of coupled particles diffusing in a heat bath in Section 4.2.

We provide a suite of appendices that are home to technical details necessary for completeness, but that would otherwise distract. Several appendices also draw out implications of information anatomy analysis. Appendix A shows that the information anatomy of a Markov system requires looking only one time step into the future and past, as expected from a similar calculation in [6]. Appendix B establishes that the causal states of a first-order Langevin equation with an analytic drift are isomorphic to the present position. Appendix C justifies why, given an infinitesimal time resolution τ, the conditional entropy of the measurement at a future time step given the present measurement can be approximated arbitrarily well by using a linearized drift term when the diffusion matrix is invertible. Appendix D then demonstrates that the entropy of the Green's function of a linear Langevin equation with a noninvertible diffusion matrix differs from that when the diffusion matrix is invertible. Finally, Appendix E applies the formulae in Appendices A–C to explore estimates of the time-local predictive information and related alternatives, used as optimization principles to choose action policies for adaptive autonomous agents [12].

Table 1. Information anatomy of first-order, n-dimensional nonlinear Langevin dynamics: $\dot{x} = -D\nabla U(x) + \eta(t)$, where $U(x)$ is analytic in x and $\eta(t)$ is zero-mean white noise with invertible diffusion matrix D, $\langle \eta(t)\eta(t')^\top \rangle = D\delta(t-t')$. Stationary distribution $\rho_{eq}(x) \propto \exp(-2U(x))$ is assumed normalizable.

Information Rates	Definition	Terms		
		$O(\tau^{-1}\log\tau)$	$O(\tau^{-1})$	$O(1)$
Stored $H_0 = C_\mu(\tau)$	$\frac{H[X_0]}{\tau}$	0	$-\int \rho_{eq}(x)\log\rho_{eq}(x)dx$	0
τ-Entropy $h_\mu(\tau)$	$\frac{H[X_0\|X_{:0}]}{\tau}$	$\frac{n}{2}$	$\log\sqrt{2\pi e\|\det D\|} + n\log\sqrt{2}$	$-\frac{1}{2}\int \nabla\cdot(D\nabla U(x))\rho_{eq}(x)dx$
Bound $b_\mu(\tau)$	$\frac{I[X_0;X_{\tau:}\|X_{:0}]}{\tau}$	0	$n\log\sqrt{2}$	$-\frac{1}{2}\int \nabla\cdot(D\nabla U(x))\rho_{eq}(x)dx$
Ephemeral $r_\mu(\tau)$	$\frac{H[X_0\|X_{:0},X_{\tau:}]}{\tau}$	$\frac{n}{2}$	$\log\sqrt{2\pi e\|\det D\|}$	0
Enigmatic $q_\mu(\tau)$	$\frac{I[X_{:0};X_0;X_{\tau:}]}{\tau}$	$-\frac{n}{2}$	$-\int \rho_{eq}(x)\log\rho_{eq}(x)dx - n\log 2 - \log\sqrt{2\pi e\|\det D\|}$	$\int \nabla\cdot(D\nabla U(x))\rho_{eq}(x)dx$
Elusive $\sigma_\mu(\tau)$	$\frac{I[X_{:0};X_{\tau:}\|X_0]}{\tau}$	0	0	0

2. Background

Let us first recall the information anatomy analysis of discrete-time, discrete-state processes introduced in [6]. The main object of study is a process $\mathcal{P}$: the list of all of a system's behaviors or realizations $\{\ldots x_{-2}, x_{-1}, x_0, x_1, \ldots\}$ and their probabilities, specified by the joint distribution $\Pr(\ldots X_{-2}, X_{-1}, X_0, X_1, \ldots)$. We denote a contiguous chain of random variables as $X_{0:L} = X_0 X_1 \cdots X_{L-1}$. We assume the process is ergodic and stationary ($\Pr(X_{0:L}) = \Pr(X_{t:L+t})$ for all $t \in \mathbb{Z}$) and the measurement symbols range over a finite alphabet: $x \in \mathcal{A}$. In this setting, the present X_0 is the random variable measured at $t = 0$, the past is the chain $X_{:0} = \ldots X_{-2}X_{-1}$ leading up the present and the future is the chain following the present $X_{1:} = X_1 X_2 \cdots$. (We suppress the infinite index in these.)

Shannon's various information quantities—entropy, conditional entropy, mutual information, and the like—when applied to time series are functions of the joint distributions $\Pr(X_{0:L})$. Importantly, they define an algebra of information measures for a given set of random variables [27]. James *et al.* [6] used this to show that the past and future partition the single-measurement entropy $H(X_0)$ into several measure-theoretic atoms. These include the ephemeral information:

$$r_\mu = H[X_0|X_{:0}, X_{1:}] \,,$$

which measures the uncertainty of the present knowing the past and future; the bound information:

$$b_\mu = I[X_0; X_{1:}|X_{:0}] \,,$$

which is the information shared between present, and future conditioned on past; and the enigmatic information:

$$q_\mu = I[X_0; X_{:0}; X_{1:}] \,,$$

which is the co-information between past, present and future.

For a stationary time series, the bound information is also the shared information between present and past conditioned on the future:

$$b_\mu = I[X_0; X_{:0}|X_{1:}].$$

One can also consider the amount of predictable information not captured by the present:

$$\sigma_\mu = I[X_{:0}; X_{1:}|X_0].$$

which is called the elusive information. It measures the amount of past-future correlation not contained in the present. It is nonzero if the process has "hidden states" and is therefore quite sensitive to how the state space is "observed" or coarse-grained.

The total information in the future predictable from the past (or *vice versa*) is the excess entropy:

$$\mathbf{E} = I[X_{:1}; X_{1:}] = b_\mu + \sigma_\mu + q_\mu \ .$$

The process's Shannon entropy rate h_μ can also be written as a sum of atoms:

$$h_\mu = H[X_0|X_{:0}] = r_\mu + b_\mu \ .$$

Thus, a portion of the information (h_μ) a process spontaneously generates is thrown away (r_μ) and a portion is actively stored (b_μ). Putting these observations together gives the information anatomy of a single measurement:

$$H[X_0] = q_\mu + 2b_\mu + r_\mu \ . \tag{1}$$

These quantities were originally defined for stationary processes, but easily carry over to a nonstationary process of finite Markov order. (See Appendix A.)

The burden of the following is to analyze the limit from the discrete-time, discrete-value processes just discussed to continuous-time, continuous-value processes. Suppose that observations are made at very small intervals of duration τ. Then, the observation at time $t_n = n\tau$ is now labeled $X_{n\tau}$, and the past $X_{:0}$ is now denoted $\ldots X_{-2\tau}X_{-\tau}$ instead of $\ldots X_{-2}X_{-1}$. Rather than entropy or mutual information per observed symbol, as in the discrete time setting, we define an entropy or mutual information per elapsed time unit; that is, informational rates. A step in this direction is to normalize the information measures defined above by the observation interval:

$$\begin{aligned}
r_\mu(\tau) &= H[X_0|X_{:0}, X_{\tau:}]/\tau \ , \\
b_\mu(\tau) &= I[X_{\tau:}; X_0|X_{:0}]/\tau \ , \\
q_\mu(\tau) &= I[X_{:0}; X_0; X_{\tau:}]/\tau \ , \\
\sigma_\mu(\tau) &= I[X_{:0}; X_{\tau:}|X_0]/\tau \ ,
\end{aligned}$$

and

$$H_0(\tau) = H[X_0]/\tau \ .$$

We normalize the entropy $H[X_0]$ of a single symbol by the time resolution τ to preserve the form of the information-theoretic relationship given in Equation (1). In doing so, we no longer interpret $H_0(\tau)$ as the entropy of a single measurement symbol, but rather as the number of bits per unit time required to encode the time series in a model-free manner.

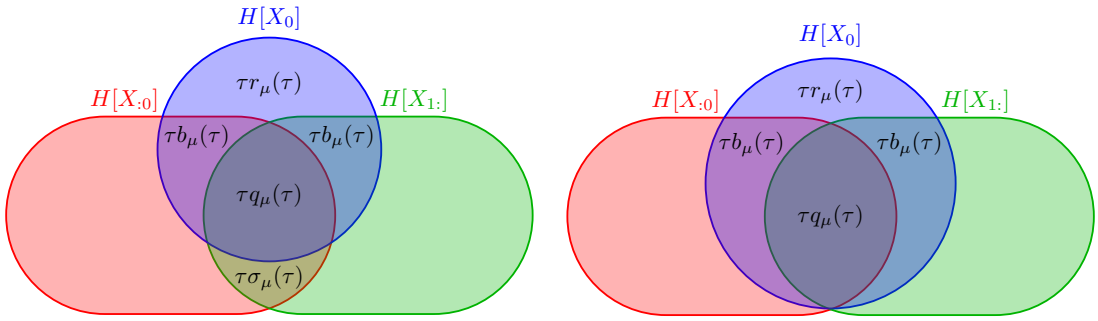
In contrast with the original information anatomy interpretation given in [6], we think of $h_\mu(\tau)$ as the minimal achievable coding rate, were we to build a maximally predictive model. In this time

normalization, terms of order τ or higher are ignored. These definitions then lead to the τ-entropy rate familiar in the discrete-time, continuous-value setting [2,25,28]:

$$h_\mu(\tau) = H[X_0|X_{:0}]/\tau \ .$$

More natural definitions of these quantities might involve a fully continuous-time development that avoids the $\log \tau$ divergences of the τ entropy rate [29]. As noted in the Introduction, however, we leave alternative developments for the future. When considering continuous-value processes, we use the differential entropy, thereby regularizing away the $\log \epsilon$ divergences seen in the (ϵ, τ)-entropy rate [25].

Figure 1. Information anatomy of a stationary continuous-time process graphically depicted using information diagrams. Although the past entropy $H[X_{:0}]$ and the future entropy $H[X_{\tau:}]$ typically are infinite, space limitations constrain us to draw them with finite areas. (**a**) Information diagram for the anatomy of a process's single observation X_0 in the context of its past $X_{:0}$ and its future $X_{\tau:}$ (after [6], with permission). (**b**) Information diagram for the anatomy of a Markov process, in which the present X_0 causally shields the past from future. The elusive information $\sigma_\mu(\tau)$ vanishes.



(a) (b)

Figures 1(a) and 1(b) give information diagrams that illustrate the algebra of the information measure atoms just defined. There, the entropy of a set is the sum of the entropy of its atoms. This reveals several useful linear dependencies that were originally noted in [6]:

$$H_0(\tau) = r_\mu(\tau) + 2b_\mu(\tau) + q_\mu(\tau) \ ,$$
$$h_\mu(\tau) = r_\mu(\tau) + b_\mu(\tau) \ ,$$

and

$$\mathbf{E}/\tau = q_\mu(\tau) + \sigma_\mu(\tau) + b_\mu(\tau) \ .$$

For a Markov process, illustrated in Figure 1b, the elusive information vanishes:

$$\sigma_\mu(\tau) = 0 \ .$$

Therefore, in this case, if we find expressions for $H_0(\tau)$, $h_\mu(\tau)$ and $b_\mu(\tau)$, then we can find $r_\mu(\tau)$, $q_\mu(\tau)$ and $\mathbf{E}/\tau$ via:

$$q_\mu(\tau) = H_0(\tau) - h_\mu(\tau) - b_\mu(\tau) \ , \quad (2)$$
$$r_\mu(\tau) = h_\mu(\tau) - b_\mu(\tau) \ , \quad (3)$$

and

$$\mathbf{E}/\tau = H_0(\tau) - h_\mu(\tau) \ . \quad (4)$$

3. Information Anatomy of Stochastic Dynamical Systems

To determine a process's information anatomy, one must calculate entropies and conditional entropies of the joint probability distribution of the entire past, the present, and the entire future. In the general case, this is challenging. However, since the first-order Langevin equations we consider are Markov, we have:

$$\tau h_\mu(\tau) = H[X_\tau|X_0] \tag{5}$$

and

$$\tau b_\mu(\tau) = H[X_\tau; X_{-\tau}] - H[X_\tau; X_0] \; . \tag{6}$$

(Appendix A provides the derivation.) Therefore, to calculate a Markov process's information anatomy, we need only the joint probability distribution of three successive measurements instead of the joint probability distribution of the present and semi-infinite past and future. To further simplify the calculation of conditional entropies, we assume that τ is small enough that the entropy of the Green's function—*i.e.*, the transition probabilities $P(x', t+\tau|x, t)$—is well approximated by the entropy of a corresponding Gaussian. This is exactly true for a linear Langevin equation. For a nonlinear Langevin equation, the Gaussian approximation is valid in the limit of infinitesimal τ. (Appendix C calculates small-τ approximations for the variance of this Gaussian.) We do not approximate the stationary distribution of a nonlinear Langevin equation by a Gaussian, however, and this means that the joint probability distribution over successive measurements is in general highly non-Gaussian. Finally, we assume that all derived information anatomy quantities are finite (at finite τ) and that there is a normalizable stationary probability distribution.

Appendix B shows that, for first-order Langevin dynamics, the single-measurement entropy $H[X_0]$ is the process's statistical complexity C_μ [3,4]. The result is that the information anatomy analysis decomposes this causal-state information into:

- that useful for prediction or retrodiction beyond the information provided by the causal states at the previous time step—the bound information b_μ;
- that useful for both prediction and retrodiction—the co-information q_μ; and
- that useless for both prediction and retrodiction—the ephemeral information rate r_μ.

This is a similar, but finer C_μ decomposition than considered in [30]. There, and more generally, $C_\mu = \mathbf{E} + \chi$. That is, the state information consists of that shared with the future ($\mathbf{E}$) and information not shared with the future, but that must be stored to implement optimal prediction—the crypticity χ [31]. Together with these observations, Equation (4) reminds us that $\chi = h_\mu$ for Markov processes, as originally noted for finite-range one-dimensional spin systems [32].

3.1. Nonlinear Langevin Dynamics

Consider an n-dimensional nonlinear Langevin equation:

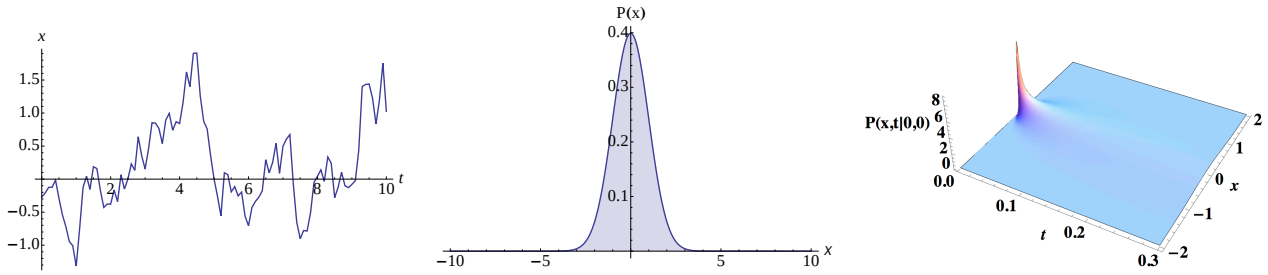
$$\frac{dx}{dt} = -D\nabla U(x) + \eta(t) \; ,$$

where $x \in \mathbb{R}^n$, $U(x)$ is an analytic potential function and $\eta(t)$ is zero-mean white noise with diffusion matrix D: $\langle \eta_i(t) \rangle = 0$ and $\langle \eta_i(t)\eta_j(t') \rangle = D_{ij}\delta(t-t')$. The diffusion coefficients $D_{ij} = D_{ji}$ are assumed to be independent of x and such that $\det D \neq 0$. The following (well known) stationary distribution is derived by converting the stochastic differential equation into its Fokker–Planck equation form:

$$\rho_{eq}(x) = \frac{1}{Z} \exp\left(-2U(x)\right) \ , \tag{7}$$

where $Z = \int e^{-2U(x)} dx$. We assume that this is the stationary probability distribution experienced by the particle and that it is normalizable: $Z < \infty$. (See Figure 2 for simulation results in one dimension.)

Figure 2. (**a**) Particle diffusing according to $\dot{x} = -x + \eta(t)$ with diffusion coefficient $D = 1$. A finite-time trajectory $x(t)$ followed by the diffusing particle. (**b**) Over infinite time, the particle experiences positions distributed according to the probability density function $\rho_{eq}(x)$ in Equation (7), calculated as a normalized histogram of particle positions. (**c**) If the previous particle position is known, a future position can be determined with less uncertainty than if no previous particle position is known. The probability $\Pr(x, t|0, 0)$ of being in position x at a time t differs from the equilibrium probability distribution $\rho_{eq}(x)$, if we know the position of the particle at a previous time; e.g., $x(0) = 0$.



(a) (b) (c)

The time-discretization normalized entropy of a measurement is:

$$H_0 = -\frac{1}{\tau} \int \rho_{eq}(x) \log \rho_{eq}(x) dx \ . \tag{8}$$

The conditional entropies $H[X_\tau|X_0]$ and $H[X_\tau|X_{-\tau}]$ in Equations (5) and (6) can be calculated, simplifying if the conditional probabilities $\Pr(X_\tau|X_0)$ and $\Pr(X_\tau|X_{-\tau})$ are Gaussians, using:

$$H[X_\tau|X_0] = \int \rho_{eq}(x') H[X_\tau|X_0 = x'] dx'$$

and

$$H[X_\tau|X_{-\tau}] = \int \rho_{eq}(x') H[X_\tau|X_{-\tau} = x'] dx' \ .$$

This yields:

$$H[X_\tau|X_0 = x'] = \tfrac{1}{2} \log(2\pi e |\det \mathrm{Var}(X_\tau)_{p(X_\tau|X_0=x')}|) \tag{9}$$

and

$$H[X_\tau|X_{-\tau} = x'] = \tfrac{1}{2} \log(2\pi e |\det \mathrm{Var}(X_\tau)_{p(X_\tau|X_{-\tau}=x')}|) \ . \tag{10}$$

Appendix C gives a plausibility proof that the conditional distributions $\Pr(X_\tau|X_0)$ and $\Pr(X_\tau|X_{-\tau})$ are Gaussian to $o(\tau)$ over a region of $\mathbb{R}^n$ with measure arbitrarily close to one. The entropies of these Gaussians are calculable to leading and subleading order in τ using a linearized version of the nonlinear Langevin equation about the initial position:

$$\frac{dx}{dt} = \nabla U(x)|_{x=x'} + A(x')(x - x') + \eta(t) + O(||x - x'||^2) \; ,$$

where $A(x')$ is a matrix with entries $(A(x'))_{ij} = \partial(D\nabla U)_j/\partial x_i$. (This is similar, but not identical to the approximation used in [12]. Appendix E comments on the differences.) From Appendix C, we have that:

$$\mathrm{Var}(X_\tau)_{p(X_\tau|X_0=x')} = D\tau + \frac{\nabla\mu(x)D + D(\nabla\mu(x))^\top}{2}\tau^2 + O(\tau^3) \tag{11}$$

and, similarly,

$$\mathrm{Var}(X_\tau)_{p(X_\tau|X_{-\tau}=x')} = 2D\tau + 2(\nabla\mu(x)D + D(\nabla\mu(x))^\top)\tau^2 + O(\tau^3) \; . \tag{12}$$

Substituting Equations (11) and (12) into Equations (9) and (10), respectively, gives, with some algebra:

$$H[X_\tau|X_{-\tau}] = -\tau \int \rho_{eq}(x)\nabla \cdot (D\nabla U(x))dx + \log\sqrt{2^{n+1}\pi e|\det D|\tau^n} \tag{13}$$

and:

$$H[X_\tau|X_0] = -\frac{\tau}{2}\int \rho_{eq}(x)\nabla \cdot (D\nabla U(x))dx + \log\sqrt{2\pi e|\det D|\tau^n} \; . \tag{14}$$

Substituting Equation (14) into Equation (5), we find that:

$$h_\mu(\tau) = \frac{n\log\sqrt{2\tau}}{\tau} + \frac{\log\sqrt{\pi e|\det D|}}{\tau} - \frac{1}{2}\int \rho_{eq}(x)\nabla \cdot (D\nabla U)dx + o(1) \; . \tag{15}$$

The leading order term is recognizable as an (ϵ, τ)-entropy rate of the Ornstein–Uhlenbeck process [25], except that the ϵ has been regularized away, since we used Shannon's differential entropy. Substituting Equations (13) and (14) into Equation (6), we find the bound information rate:

$$b_\mu(\tau) \;\; = \;\; \frac{n\log\sqrt{2}}{\tau} - \frac{1}{2}\int \rho_{eq}(x)\nabla \cdot (D\nabla U(x))dx + o(1) \; . \tag{16}$$

Thus, the rate of active information storage depends on the dimension of the state space to leading order in τ, but its nondivergent part depends on the average curvature of the potential.

From these quantities, all other anatomy measures follow. Substituting Equations (15) and (16) into Equation (3), we find that the ephemeral information is:

$$r_\mu(\tau) = \frac{n\log\sqrt{\tau}}{\tau} + \frac{\log\sqrt{2\pi e|\det D|}}{\tau} + o(1) \; . \tag{17}$$

Unsurprisingly, the dissipated information—that entropy created in the present useful for neither predicting nor retrodicting—depends only on the noisiness of the dynamics and not on the drift.

Finally, the enigmatic information—that shared between past, future, and present—follows by substituting Equations (8)–(16) into Equation (2):

$$q_\mu(\tau) = -\frac{1}{\tau}\int \rho_{eq}(x)\log\rho_{eq}(x)dx - \frac{n\log(2\sqrt{\tau})}{\tau} - \frac{\log\sqrt{\pi e|\det D|}}{\tau} - \int \rho_{eq}(x)\nabla\cdot(D\nabla U)dx + O(\tau) \ .$$

It is interesting to consider how q_μ changes as the stochasticity of the system increases: the stationary distribution $\rho_{eq}(x)$ flattens out, leading to an unbounded increase in H_0. This is counteracted by an unbounded increase in the entropy rate.

We can also bound the bound information rate when ∇U grows more slowly than e^{-2U} with $||x||$. Then, integration by parts applied to Equation (16) gives:

$$b_\mu(\tau) = \frac{n\log\sqrt{2}}{\tau} - \frac{1}{2}\int (\nabla U)^\top D(\nabla U)\ \rho_{eq}(x)dx \ .$$

When D is positive semidefinite, with $D = v^\top v$ for some vector v, then:

$$b_\mu(\tau) \leq \frac{n\log\sqrt{2}}{\tau} \ .$$

Therefore, $b_\mu(\tau)$ is maximized when the potential well is as flat as possible, while maintaining $Z < \infty$.

3.2. Linear Langevin Equation with Noninvertible Diffusion

What if the invertibility of the diffusion matrix is relaxed? In particular, do we still have qualitatively the same information anatomy if a subsystem of the stochastic dynamical system evolves deterministically? How does this affect the information generation and storage properties? To this end, suppose $x = (x_d \quad x_n)^\top$ with $x \in \mathbb{R}^k$ and $m = \dim(x_d)$, where x_d evolves deterministically and x_n stochastically:

$$\frac{dx_d}{dt} = A_d + B_{dd}x_d + B_{dn}x_n \tag{18}$$

$$\frac{dx_n}{dt} = A_n + B_{nd}x_d + B_{nn}x_n + \eta(t) \ . \tag{19}$$

Again, $\eta(t)$ is white noise with $\langle\eta(t)\rangle = 0$ and $\langle\eta(t)\eta(t')^\top\rangle = D\delta(t-t')$, where D is invertible. Taken together, though, this is a linear Langevin equation for x with a noninvertible diffusion matrix. Naively assuming that the deterministic subsystem evolves with a small amount of noise, Equation (16) would apply and give, for example, to $O(\tau)$:

$$b_\mu = \frac{(n+m)\log 2}{2\tau} + \frac{\mathrm{tr}(B_{dd}) + \mathrm{tr}(B_{nn})}{2} \ .$$

However, this assumption is incorrect; the noiseless limit is singular.

Since Equations (18) and (19) specify a linear Langevin equation for x, its Green's function is Gaussian. For simplicity's sake, we assume that $B_{dn}D_{nn}B_{dn}^\top$ is invertible, though it is certainly possible to derive more complicated expressions for information anatomy quantities if this does not hold. From Appendix D, to $O(\tau)$ the entropy rate is:

$$h_\mu(\tau) = \frac{(n+3m)\log\tau}{2\tau} - \frac{m\log\sqrt{12}}{\tau} + \frac{\log\sqrt{2\pi e|\det D_{nn}||\det B_{dn}D_{nn}B_{dn}^\top|}}{\tau} + \frac{\mathrm{tr}(B_{dd}) + \mathrm{tr}(B_{nn})}{2}$$

and the bound information is:

$$b_\mu(\tau) = \frac{(n+3m)\log 2}{2\tau} + \frac{\mathrm{tr}(B_{dd}) + \mathrm{tr}(B_{nn})}{2} \ . \tag{20}$$

Applying Equation (3), the ephemeral information rate is to $O(\tau)$:

$$r_\mu(\tau) = \frac{n+3m}{2}\frac{\log(\tau/2)}{\tau} - \frac{m\log\sqrt{12}}{\tau} + \frac{\log\sqrt{2\pi e|\det D_{nn}||\det B_{dn}D_{nn}B_{dn}^\top|}}{\tau} \ . \tag{21}$$

These answers are very different from those derived assuming that x's deterministic subsystem x_d evolves with an infinitesimal amount of noise. The bound information in Equation (20) differs from that found from naive application of Equation (16), because the pre-factor for the $\log 2/\tau$ divergence is $(n+m)/2+m$ rather than $(n+m)/2$. That is, the difference counts the dimension m of the deterministically evolving state space x_d. Thus, the deterministic subsystem allows for the active storage of more of the spontaneously generated stochasticity.

The ephemeral information in Equation (21) differs from a naive application of Equation (17) in two new ways. First, the expression in Equation (21) has an additional $O(1/\tau)$ factor that is linearly proportional to the dimension m of the deterministic subsystem. Second, the term $\log(2\pi e|\det D_{nn}||\det B_{dn}D_{nn}B_{dn}^\top|)$ can be interpreted by supposing that $B_{dn}D_{nn}B_{dn}^\top$ is the effective diffusion matrix felt by the deterministically evolving states.

These information anatomy quantities are therefore sensitive to the process's underlying noise architecture.

4. Examples

To illustrate how the information measures are helpful and interesting summaries of nonlinear Langevin dynamics, let us consider several examples.

4.1. Stochastic Gradient Descent in One Dimension

Consider a first-order nonlinear Langevin dynamics for $x \in \mathbb{R}$ in which:

$$\frac{dx}{dt} = -\frac{dU(x)}{dx} + \eta(t) \ ,$$

where $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = 2D\delta(t-t')$. The stationary distribution is:

$$\rho_{eq}(x) = \frac{1}{Z} e^{-U(x)/D} \ ,$$

with Z a normalization factor:

$$Z = \int_{-\infty}^{\infty} e^{-U(x)/D} dx \ .$$

We require that $Z < \infty$.

This process's elusive information is zero, and the ephemeral information rate is the strength of the noise. However, the bound information is:

$$b_\mu(\tau) = \frac{\log\sqrt{2}}{\tau} - \frac{1}{2Z}\int_{-\infty}^{\infty} e^{-U(x)/D}\frac{d^2U(x)}{dx^2} dx \ . \tag{22}$$

Using integration by parts, this can be rewritten:

$$b_\mu(\tau) = \frac{\log\sqrt{2}}{\tau} - \frac{1}{2D}\int_{-\infty}^{\infty}\left(\frac{dU}{dx}\right)^2 \frac{e^{-U(x)/D}}{Z}dx \ .$$

Therefore, b_μ is sensitive to the average curvature of the potential or, equivalently, to the average squared drift normalized by the diffusion constant.

In the deterministic limit, this expression simplifies. Suppose that $\{x_1^*, ..., x_m^*\}$ are the global minima of the potential function: $U(x_i^*) = \min_x U(x)$, for $i = 1, ..., m$. It follows that $\lim_{D\to 0} e^{-U(x)/D}/Z = \sum_{i=1}^m \delta(x - x_i^*)/m$. Applying this limit to Equation (22), we have:

$$\lim_{D\to 0} b_\mu(\tau) = \frac{\log\sqrt{2}}{\tau} - \frac{1}{2m}\sum_{i=1}^{m}\frac{d^2U(x)}{dx^2}|_{x=x_i^*} \ .$$

This limit is a little strange. If $D = 0$ exactly, so that we have deterministic gradient descent, then the stationary time series consists of a single measurement. The information anatomy becomes rather trivial. There is no uncertainty in the present measurement, and the past, present, and future share no information. If D is nonzero, no matter how small, however, then there is finite uncertainty in a measurement, and the past, present and future share information with one another.

As a concrete example, consider the canonical form for the cusp catastrophe [33]:

$$\frac{dx}{dt} = h + rx - x^3 + \eta(t) \ ,$$

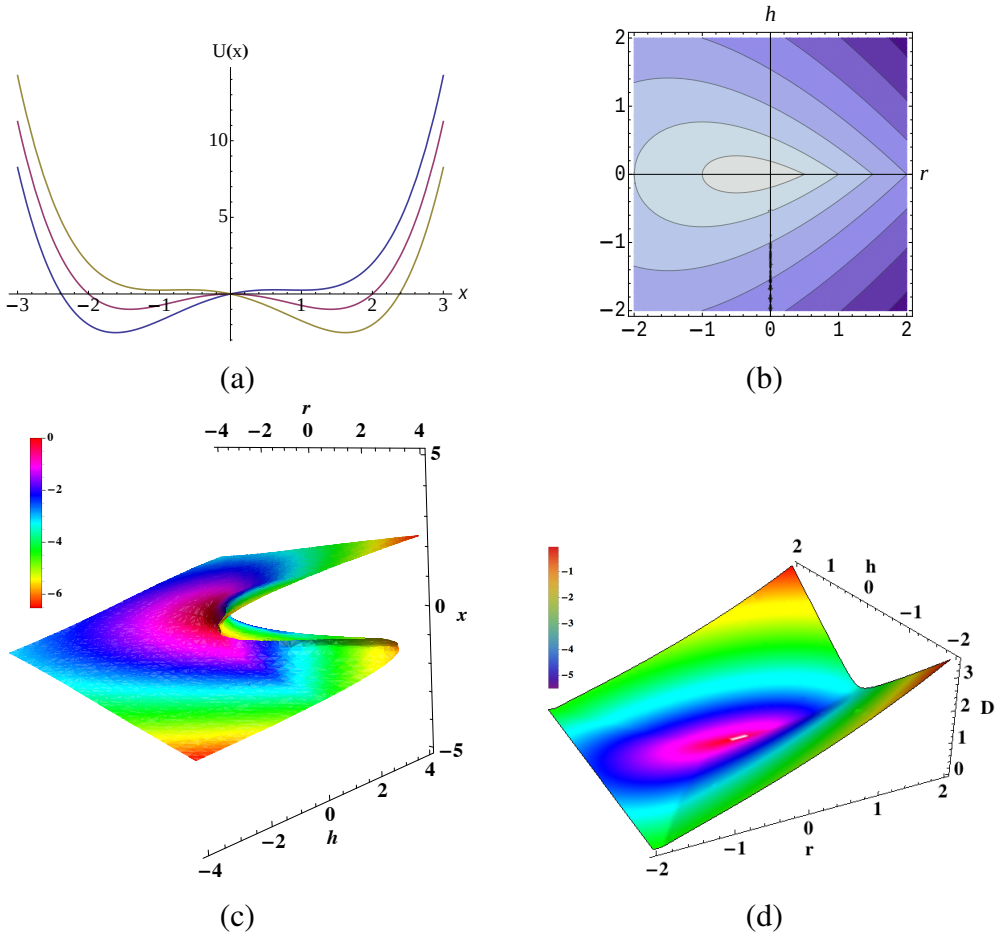
with additive noise where $\langle \eta(t)\eta(t')\rangle = 2D\delta(t - t')$. The potential function is $U(x) = \frac{1}{4}x^4 - \frac{1}{2}rx^2 - hx$, and the corresponding bound information in the noiseless limit is:

$$\lim_{D\to 0} b_\mu(\tau, r, h) = \frac{\log\sqrt{2}}{\tau} + \frac{r - 3(x^*(r,h))^2}{2} \ .$$

The global minimum $x^*(r, h)$ is not everywhere differentiable in r and h, and this appears also in $b_\mu(\tau, r, h)$. See Figure 3. The contour of nondifferentiability is $h = 0$ for $r > 0$. Along the contour, the potential is symmetric, there are suddenly two global minima of $U(x)$ with $x_1^* = -x_2^*$, and so, the sign of x^* changes discontinuously across $h = 0$.

Interestingly, for double-well potentials and asymmetric single-well potentials, $b_\mu(\tau)$ is maximized at a nonzero noise level $D > 0$. This is counterintuitive: adding noise only serves to decrease the process's predictability. However, adding noise in the present affects the future in a way that cannot be predicted from the past. Since $b_\mu(\tau)$ measures the amount of information shared between the present and future not shared with the past, there is a level of stochasticity that maximizes $b_\mu(\tau)$ for some values of r and h. This is shown in Figure 3c.

Figure 3. Information anatomy of the stochastic cusp catastrophe: (**a**) Shifting from a double-well to single-well potentials as r and h are varied. Example potentials $U(x)$ for various r and h: blue/dark line, $r = 2$ and $h = -1$; purple/medium line, $r = 2$ and $h = 0$; and yellow/light line, $r = 2$ and $h = 1$. (**b**) Contour plot of the system-dependent part of the bound information rate $b_\mu(\tau)$ as a function of r and h, highlighting the global minimum x^* changing discontinuously as h moves through zero. $\lim_{D\to 0} b_\mu(\tau) - \tau^{-1} \log \sqrt{2}$ as a function of r and h: $b_\mu(\tau)$ is nondifferentiable with respect to h along $h = 0$ when $r \geq 0$. (**c**) Bound information $b_\mu(\tau)$ as it varies over the cusp catastrophe equilibria surface: Height gives the fixed points as a function of r and h. Color hue is proportional to the deterministic limit $\lim_{D\to 0} b_\mu(\tau) - \tau^{-1} \log \sqrt{2}$ at each r and h. (**d**) The bound information rate is maximized at nonzero stochasticity D for double-well potentials and asymmetric single-well potentials. D maximizing $b_\mu(\tau) - \tau^{-1} \log \sqrt{2}$ as a function of r and h: the surface is colored by $b_\mu(\tau) - \tau^{-1} \log \sqrt{2}$ at that value of D.



4.2. Particles Diffusing in a Heat Bath

Suppose N particles with positions $x_1, ..., x_N$ and masses $m_1, ..., m_N$ diffuse according to the potential function $U(x_1, ..., x_N)$ in a heat bath of temperature T. Let **x** denote the vector of concatenated

particle positions. When the inertial terms $m_i d^2 x_i / dt^2$ are negligible, an overdamped Langevin equation can be used to approximate the particles' trajectories:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \frac{1}{\gamma} M^{-1} \nabla U(\mathbf{x}) + \eta(t) \\ \langle \eta_i(t) \rangle &= 0 \\ \langle \eta_i(t) \eta_j(t') \rangle &= \frac{2 k_B T}{\gamma m_i} \delta_{i,j} \delta(t - t'). \end{aligned}$$

M is a diagonal matrix whose entries are the particle masses, and the parameter γ is a friction coefficient that controls how strongly the particles couple to the heat bath. The stationary distribution of positions x is the Boltzmann distribution:

$$\rho_{eq}(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{U(\mathbf{x})}{k_B T} \right) ,$$

where Z is the partition function:

$$Z = \int \exp\left(-\frac{U(\mathbf{x})}{k_B T} \right) d\mathbf{x} .$$

From Equation (8), the normalized single-measurement entropy is:

$$H_0 = \frac{1}{\tau} \left(\frac{\langle U(\mathbf{x}) \rangle}{k_B T} + \ln Z \right) .$$

where:

$$\langle U(\mathbf{x}) \rangle = \int U(\mathbf{x}) \frac{e^{-U(\mathbf{x})/k_B T}}{Z} d\mathbf{x} ,$$

which is simply proportional to the familiar definition of entropy in physics.

For notational ease, let $\bar{m}$ denote the geometric mean of the masses:

$$\bar{m} = \left(\prod_{i=1}^{N} m_i \right)^{1/N} ,$$

k_i the effective "spring constant" for the i^{th} particle:

$$k_i = \int \frac{\partial^2 U(\mathbf{x})}{\partial x_i^2} \frac{e^{-U(\mathbf{x})/k_B T}}{Z} d\mathbf{x} ,$$

and ω_i the effective "oscillation frequency" for the i^{th} particle:

$$\omega_i = \sqrt{k_i / m_i} .$$

From Equation (15), the entropy rate is:

$$h_\mu(\tau) = \frac{N \log \sqrt{4 k_B T \tau / \gamma \bar{m}}}{\tau} + \frac{\log \sqrt{\pi e}}{\tau} - \frac{1}{2\gamma} \sum_{i=1}^{N} \omega_i^2 + o(1) .$$

From Equation (16), the bound information is to similar order:

$$b_\mu(\tau) = \frac{N \log \sqrt{2}}{\tau} - \frac{1}{2\gamma} \sum_{i=1}^{N} \omega_i^2 + o(1) \ .$$

From Equation (17), the ephemeral information rate is:

$$r_\mu(\tau) = \frac{N \log \sqrt{2 k_B T \tau / \gamma \bar{m}}}{\tau} + \frac{\log \sqrt{2\pi e}}{\tau} + o(1) \ .$$

Several information measures appear dimensionally incorrect. This is a perennial concern when calculating the differential entropy of random variables that themselves have units. The probability density over those variables also has a dimension, and this leads to differential entropies that involve the log of a value with dimension. Implicitly, however, we chose a standard unit system, such that all quantities are dimensionless.

All of these quantities are extensive in N. The normalized entropy per measurement H_0 is proportional to the Boltzmann entropy by a factor of k_B/τ. The entropy rate $h_\mu(\tau)$ and ephemeral information $r_\mu(\tau)$ increase logarithmically with the mean squared velocity $\sqrt{\langle v^2 \rangle} = k_B T / m$. The bound information $b_\mu(\tau)$ increases when there is a larger γ. That is, it increases when there is stronger coupling between the particles and the heat bath or when there is a smaller average oscillation frequency $\sum_{i=1}^{N} \omega_i^2$. Since $\gamma \geq 0$ and $\omega_i^2 \geq 0$, the bound information is bounded above by $b_\mu(\tau) \leq \tau^{-1} N \log \sqrt{2} + O(\tau)$. To achieve this upper bound, the potential $U(\mathbf{x})$ must be "flattened out" to decrease k_i, as described in Section 3.

There are alternative models for coupled particles diffusing in a heat bath, and there is no guarantee that even the qualitative conclusions here hold true when particle trajectories are modeled according to a second-order Langevin equation, for instance.

5. Conclusions

Our calculations led to general formulae for the information anatomy of stochastic equilibria in simple, familiar systems when the time discretization was very small. We considered a first-order nonlinear Langevin equation with a normalizable stationary distribution, invertible diffusion matrix, and analytic drift. We do not expect the expressions in Section 3 to hold for larger time discretizations, though Gaussian approximations could be used to upper bound conditional entropies more generally. We also considered first-order linear Langevin equations with normalizable stationary distribution and a noninvertible diffusion matrix in Section 3.2.

An important technical consideration is that the information anatomy of Langevin stochastic dynamics is likely not unique, just as the pre-factors for the (ϵ, τ)-entropy rate of an Ornstein–Uhlenbeck process depend on definition and approximation procedure [25,28]. However, further calculations give us reason to believe that the qualitative scaling seen with drift and diffusion holds regardless of the approximation method. This parallels the way that the (ϵ, τ)-entropy rate estimates for an Ornstein–Uhlenbeck process all increase with the diffusion coefficient. That said, a complete understanding of how information anatomy estimates vary with technique requires further study; alternatives to which the Introduction alluded. We hope that our results are sufficiently compelling to motivate further efforts.

With this caveat in mind, let us focus on qualitative rather than quantitative conclusions. Even though the entropy rate is typically viewed as a measure of randomness, some of that randomness is useful for prediction, that is, the bound information (shared between present and future, but not contained in the past), and we showed that it is sensitive to drift and the diffusion matrix. In contrast, we showed that the ephemeral information—information in the present useless for predicting or retrodicting—is sensitive only to the diffusion and not the drift. In short, for stochastic equilibria, the entropy rate consists of a quantity (ephemeral information) that has to do with a process's inherent noisiness and a quantity (bound information) that has only to do with the underlying process regularities.

A key lesson is that information anatomy measures are sensitive to process organization. Section 3.2 showed that the divergent components of the information anatomy of linear Langevin dynamics changes discontinuously whenever one of the diffusion coefficients vanishes. This sensitivity to underlying process structure could also be a feature rather than a defect. For instance, if we know that the underlying process is a first-order linear Langevin equation, then one could infer the dimension of the deterministically evolving state space by comparing known τ-scaling relations in Section 3 with empirically determined scaling relations.

This brings us to discuss what was learned from the several example applications. Section 4.1 showed that the bound information picks up different features than one finds in a bifurcation diagram. In the noiseless limit, the cusp catastrophe b_μ is nondifferentiable on the line $h = 0$ for $r \geq 0$, because the location of the global minimum of the potential function changes discontinuously across that contour. Moreover, this is not related to the bifurcation contour $h = \pm 2r^{3/2}/3\sqrt{3}$ [33] where the number of equilibria changes from two to one or *vice versa*, which has no apparent signature in the bound information. However, in these calculations, we did not avoid the "ultraviolet catastrophe". We embraced it, since we could then evaluate the information anatomy for general nonlinear Langevin equations by linearizing. If one evaluates the information anatomies of these types of stochastic dynamics when the time discretization is not infinitesimal, however, then signatures of bifurcations should show up in the bound information as they do for the finite-time predictable information or excess entropy [16,34].

Section 4.2 calculated the information anatomy of coupled particles in a heat bath. Historically, statistical physics has been primarily concerned with H_0, the entropy of a single measurement symbol, since its changes are proportional to heat loss [35]. However, the point of this example is that alternative information-theoretic quantities capture other behavioral properties of particles diffusing in a heat bath. As an application of this analysis, it will be worth exploring how the information anatomy measures reflect the trade-off between stable information storage and heat loss in the context of Maxwell-like demons [36].

To close our discussion of applications, we briefly mention the use of information measures to express optimization principles that guide adaptive agents. A Markov process's bound information has been used as an optimization measure called the time-local predictive information (TiPi) [12]. Moreover, the class of systems used there and for which TiPi was calculated are exactly the first-order nonlinear Langevin dynamics analyzed here. Due to the similarities in setup and approach, Appendix E compares alternative TiPi measures. Generally, an agent that wishes to maximize its TiPi will be driven into unstable regions of the potential landscape on which it diffuses. However, Appendix E shows that the similarly motivated, but alternative, optimization measures lead to different adaptive strategies. More

investigation is required to compare such strategies to those seen in biological agents before general principles of adaptive behavior will be understood.

Acknowledgments

The authors are indebted to one of the anonymous reviewers for a particularly detailed critique and also for pointing them to Girsanov transformations. The authors thank the Santa Fe Institute for its hospitality during visits. J.P.C. is a Santa Fe Institute External Faculty member. This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-12-1-0234. S.M. was funded by a National Science Foundation Graduate Student Research Fellowship and the U.C. Berkeley Chancellor's Fellowship.

Author Contributions

Both authors contributed equally to conception and writing. S.M. performed the bulk of the calculations and numerical computation. Both authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

A. Information Anatomy of a Markov Process

If the system at hand is Markov, then the information anatomy simplifies tremendously since one need only consider single time steps into the future and into the past. As a result, many of the Markov formulae are special cases of those developed in [6] for more complex processes, but are derived here for completeness.

For notational ease, we use the discrete-time notation in which $X_{t:t'}$ is the random variable of measurements $X_t, X_{t+1}, ..., X_{t'-1}$. For a Markov process the immediately preceding observation "shields" the future from the past:

$$\Pr(X_n = x_n | X_{-m:n} = x_{-m:n}) = \Pr(X_n = x_n | X_{n-1} = x_{n-1}) \ .$$

Additionally, it becomes relatively easy to calculate the information anatomy measures, since the sequence probabilities simplify:

$$\Pr(X_{-m:n+1} = x_{-m:n+1}) = \Pr(X_{-m} = x_{-m}) \prod_{k=-m}^{n-1} \Pr(X_{k+1} = x_{k+1} | X_k = x_k) \ .$$

For example, the entropy rate becomes:

$$\begin{aligned} h_\mu &= H[X_0 | X_{:0}] \\ &= H[X_0 | X_{-1}] \ . \end{aligned}$$

Moreover, all information shared between the past and future goes through the present:

$$\begin{aligned}\sigma_\mu &= I[X_{:0}; X_{1:}|X_0] \\ &= H[X_{1:}|X_0] - H[X_{1:}|X_{:1}] \\ &= H[X_{1:}|X_0] - H[X_{1:}|X_0] \\ &= 0 \ .\end{aligned}$$

Finally, the mutual information between the present and the future conditioned on the past (bound information) is:

$$\begin{aligned}b_\mu &= I[X_0; X_{1:}|X_{:0}] \\ &= H[X_{1:}|X_{:0}] - H[X_{1:}|X_{:1}] \\ &= H[X_1|X_{:0}] - H[X_1|X_{:1}] + H[X_{2:}|X_{:0}, X_1] - H[X_{2:}|X_{:2}] \\ &= H[X_1|X_{-1}] - H[X_1|X_0] \\ &= I[X_1; X_0|X_{-1}] \ .\end{aligned}$$

This equality is evident from the information diagram of Figure 1b. The other information anatomy measures follow from b_μ and h_μ via identities given in Section 2:

$$\begin{aligned}r_\mu &= h_\mu - b_\mu \\ &= H[X_0|X_{-1}] - I[X_1; X_0|X_{-1}]\end{aligned}$$

and

$$\begin{aligned}q_\mu &= H[X_0] - h_\mu - b_\mu \\ &= H[X_0] - H[X_0|X_{-1}] - I[X_1; X_0|X_{-1}] \ .\end{aligned}$$

The excess entropy follows as the sum:

$$\begin{aligned}\mathbf{E} &= \sigma_\mu + q_\mu + b_\mu \\ &= H[X_0] - h_\mu \\ &= H[X_0] - H[X_0|X_{-1}] \ .\end{aligned}$$

As stated in Section 2, to normalize these measures as rates (entropies per unit time rather than per measurement), we simply divide the above by the time discretization τ:

$$\begin{aligned}h_\mu(\tau) &= \frac{H[X_0|X_{-\tau}]}{\tau} \\ b_\mu(\tau) &= \frac{I[X_\tau; X_0|X_{-\tau}]}{\tau} \\ r_\mu(\tau) &= \frac{H[X_0|X_{-\tau}] - I[X_\tau; X_0|X_{-\tau}]}{\tau} \\ q_\mu(\tau) &= \frac{H[X_0] - H[X_0|X_{-\tau}] - I[X_\tau; X_0|X_{-\tau}]}{\tau} \ .\end{aligned}$$

If the system is Markov, one only needs the joint distribution of three successive measurements to calculate the information anatomy. Thus, the formulae derived here also can be used as time-local measures for nonstationary dynamics despite the subtleties of defining a measure over bi-infinite time series in general [37]. Similar manipulations can be applied more generally to find the information anatomy of finite-order Markov processes.

B. Statistical Complexity is the Entropy of a Measurement

The statistical complexity C_μ is the entropy of the probability distribution over causal states. Causal states themselves are groupings of pasts that are partitioned according to the predictive equivalence relation $\sim_\epsilon$ [4]:

$$x_{:0} \sim_\epsilon x'_{:0} \Leftrightarrow \Pr(X_{0:}|X_{:0} = x_{:0}) = \Pr(X_{0:}|X_{:0} = x'_{:0}) \ .$$

Although causal states can be difficult to determine for complex processes, they are particularly easy for Markov processes (and finite-order Markov processes). Recall that a Markov process is defined by single-time step shielding:

$$\Pr(X_{0:}|X_{:0}) = \Pr(X_0|X_{-\tau}) \Pr(X_{1:}|X_0) \ .$$

It follows that:

$$\Pr(X_{0:}|X_{:0} = x_{:0}) = \Pr(X_{0:}|X_{:0} = x'_{:0}) \Leftrightarrow \Pr(X_0|X_{-1} = x_{-1}) = \Pr(X_0|X_{-1} = x'_{-1}) \ .$$

Therefore, for a Markov process, groupings of pasts in which only the last measurement is recorded constitute at least a prescient partition. If:

$$\Pr(X_0|X_{-1} = x) = \Pr(X_0|X_{-1} = x') \Leftrightarrow x = x' \ ,$$

then we can conclude that the causal states are simply groupings of pasts with the same last measurement: $\epsilon(x_{:0}) = x_{-1}$. In that case, the causal state space $\mathcal{S}$ is isomorphic to the alphabet of the process $\mathcal{A}$ and the statistical complexity is the entropy of a single measurement: $C_\mu = H[X_0]$.

First-order Langevin equations generate Markov time series. Our claim, then, is that the stochastic differential equations considered here produce time series for which:

$$\Pr(X_0|X_{-\tau} = x) = \Pr(X_0|X_{-\tau} = x') \Leftrightarrow x = x' \ .$$

Therefore, the causal states are isomorphic to the present measurement X_0, and the statistical complexity is $C_\mu = H[X_0]$. Implicit in these calculations is an assumption that the transition probabilities $\Pr(X_0|X_{-\tau})$ for a given stochastic differential equation exist and are unique, which is satisfied, since the drift term is analytic [38].

For intuition, consider linear Langevin dynamics for an Ornstein–Uhlenbeck process:

$$dX_t = Adt + BX_t dt + \sqrt{D} dW_t \ .$$

As described in Appendix D and many other places (e.g., [21]), the transition probability density $\Pr(X_t|X_0 = x)$ is a Gaussian:

$$\Pr(X_t|X_0 = x) \sim \mathcal{N}\left(e^{Bt}x + e^{Bt}\int_0^t e^{-Bt'}Adt', \int_0^t e^{Bt'}De^{B^\top t'}dt'\right) .$$

For $\Pr(X_t|X_0 = x) = \Pr(X_t|X_0 = x')$, the means and variances of the above probability distribution must match, meaning that $e^{Bt}x = e^{Bt}x' \Rightarrow x = x'$. Therefore, for an Ornstein–Uhlenbeck process, the causal states are indeed isomorphic to the present measurement and the statistical complexity is $H[X_0]$. The key here is that although $\Pr(X_t|X_0 = x)$ may quickly forget its initial condition x, for any finite-time discretization, the transition probability $\Pr(X_t|X_0 = x)$ still depends on x.

In the more general case, we have a nonlinear Langevin equation:

$$dX_t = -D\nabla U dt + \sqrt{2D}dW_t ,$$

where the stationary distribution ρ_{eq} exists and is normalizable. Our goal is to show that if $\Pr(X_t|X_0 = x) = \Pr(X_t|X_0 = x')$, then $x = x'$. The transition probability $\Pr(X_t = x|X_0 = x')$ is a solution to the corresponding Fokker–Planck equation:

$$\frac{\partial\rho(x,t)}{\partial t} = -\nabla\cdot(\mu(x)\rho(x,t)) + D\nabla^2\rho(x,t) ,$$

with initial condition $\rho(x,0) = \delta(x - x')$. As in [38], we can use an eigenfunction expansion to show that $\rho(x,t|x',0)$ cannot equal $\rho(x,t|x'',0)$ unless $x' = x''$ for finite time t. Therefore, $\Pr(X_t|X_0 = x') = \Pr(X_t|X_0 = x'') \Rightarrow x' = x''$. This implies that the causal states are again isomorphic to the present measurement and the statistical complexity is $C_\mu = H[X_0]$.

To summarize, this application of computational mechanics [3,4] to Langevin stochastic dynamics shows that the entropy of a single measurement is also the process's statistical complexity C_μ. Recall that the latter is the entropy of the probability distribution over the causal states which, in turn, are groupings of pasts that lead to equivalent predictions of future behavior. Therefore, for the stochastic differential equations considered here, causal states simply track the last measured position.

What the information anatomy analysis reveals, then, is that not all of the information required for optimal prediction is predictable information about the future. In other words, Langevin stochastic dynamics are inherently cryptic [30,31]. Unfortunately, as is so often the case, the necessary and the apparent come packaged together and cannot be teased apart without effort.

C. Approximating the Short-Time Propagator Entropy

The study of stochastic differential equations and short-time propagator approximations is mathematically rich and, as noted in the introduction, the application to nonlinear diffusion has a long history [21]. What follows is a brief sketch, not a rigorous proof, that glosses over important pathological cases.

Consider the nonlinear Langevin equation:

$$\frac{dx}{dt} = -D\nabla U(x) + \eta(t) , \tag{A1}$$

with driving noise satisfying $\langle \eta(t) \rangle = 0$ and $\langle \eta(t) \eta^\top(t') \rangle = D\delta(t - t')$, where $\det D \neq 0$. Let $p(x|x')$ be the transition probability $\Pr(X_t = x | X_0 = x')$ for the system in Equation (A1). From arguments in [38], it exists and is uniquely defined. Let $q(x|x')$ be a Gaussian with the same mean and variance as $p(x'|x)$.

We show that $H[p] = H[q] + o(\tau)$ where $H[p] = -\int p(x|x') \log p(x|x') dx$ and $H[q] = -\int q(x|x') \log q(x|x') dx$. Note that here, and in the following, we suppress notation for the dependence of these quantities on x', using the shorthand $H[p] \equiv H[X|X' = x']$ and the like. First, consider:

$$\begin{aligned} D_{KL}[p||q] &= \int p(x|x') \log \frac{p(x|x')}{q(x|x')} dx \\ &= \int p(x|x') \log p(x|x') dx - \int p(x|x') \log q(x|x') dx \\ &= -H[p] - \int p(x|x') \log q(x|x') dx \ . \end{aligned}$$

Since $q(x|x')$ is the maximum entropy distribution consistent with the mean and the variance of $p(x|x')$, averages of $\log q(x|x')$ with respect to p are the same as those with respect to q. Specifically, if $\bar{x}$ is the mean:

$$\bar{x} = \int x p(x|x') dx = \int x q(x|x') dx$$

and if $C(x')$ is the variance:

$$\begin{aligned} C(x') &= \int (x - \bar{x})(x - \bar{x})^\top q(x|x') dx \\ &= \int (x - \bar{x})(x - \bar{x})^\top p(x|x') dx \ , \end{aligned}$$

then q is the normal distribution consistent with that mean and variance:

$$q(x|x') = \frac{1}{\sqrt{2\pi |\det C(x')|}} \exp\left(-\frac{1}{2} (x - \bar{x})^\top C(x')^{-1} (x - \bar{x}) \right) \ .$$

From this, we derive:

$$\begin{aligned} \int p(x|x') \log q(x|x') dx &= \int p(x|x') \log \frac{e^{-\frac{1}{2}(x-\bar{x})^\top C(x')^{-1}(x-\bar{x})}}{\sqrt{2\pi |\det C(x')|}} \\ &= -\frac{1}{2} \int (x - \bar{x})^\top C(x')^{-1} (x - \bar{x})) p(x|x') dx - \log \sqrt{2\pi |\det C(x')|} \ . \end{aligned}$$

Since the mean and variance for p and q are consistent, we have:

$$\begin{aligned} \int p(x|x') \log q(x|x') dx &= -\frac{1}{2} \int (x - \bar{x})^\top C(x')^{-1} (x - \bar{x})) q(x|x') dx - \log \sqrt{2\pi |\det C(x')|} \\ &= \int q(x|x') \log q(x|x') dx \\ &= -H[q] \end{aligned}$$

and, thus:

$$D_{KL}[p||q] = H[q] - H[p] \ .$$

We wish to show that $D_{KL}[p||q]$ is at least of $o(\tau)$. Then, we also want to show that $H[q]$ can be determined to $o(\tau)$ from the linearized Langevin equation:

$$\frac{dx}{dt} = \mu(x') + \frac{\partial \mu(x)}{\partial x}|_{x=x'}(x - x') + \eta(t) \; .$$

Then, we would be able to approximate $H[p]$ to $o(\tau)$ by $H[q^{linearized}]$, where $q^{linearized}$ is the transition probability that results when we locally linearize the drift.

Our strategy is to construct a series expansion for the moments of p in the timescale τ, as in [39]. Immediately, with that statement, we run into a problem. Moments do not uniquely specify a distribution unless an additional condition (e.g., Carleman's condition) is satisfied. We will address this issue at the end of this Appendix. The second issue we find is that the sum of higher-order terms in the moment expansion is often divergent, but we have circumvented this limitation by working with infinitesimal time discretizations.

The Kullback–Leibler divergence is invariant to changes in the coordinate system and, for reasons that become apparent later, it is useful to move to the parametrization $z = (x - \bar{x})/\sqrt{t}$. In a slight abuse of notation, $p(z|x')$ and $q(z|x')$ will be used to denote the re-parametrized distributions $p(x|x')$ and $q(x|x')$. Our moment expansion will show that all moments of $p(z|x')$ and $q(z|x')$ differ by a quantity that is at most of $O(\tau^{3/2})$, which implies that $p(z|x') = q(z|x') + \tau^{3/2}\delta q$, where δq is at most of $O(1)$ in τ. From that, it would follow that $D_{KL}[q + \tau^{3/2}\delta q||q] = (\tau^{3/2})^2 \mathcal{I}[q]$, where $\mathcal{I}[q]$ is the Fisher information of a Gaussian (and hence bounded) and that $H[p] = H[q]$ to $O(\tau^3)$. That same moment expansion will show that the covariance and mean of p differ from the covariance and mean of $q^{linearized}$ by a correction term of at most $O(\tau^2)$. From this, it follows that $H[q]$ is $H[q^{linearized}]$ to $o(\tau)$. The bottleneck in this approximation scheme is not approximating the transition probability as a Gaussian, but rather approximating the covariance of that Gaussian by the covariance of the locally linearized stochastic differential equation.

For intuition and simplicity, we start with the one-dimensional example. This is similar in flavor to the approach in [39], but our point differs: we wish to understand how well we can approximate the full system with a linearized drift term. The stochastic differential equation for $x \in \mathbb{R}$ is:

$$\frac{dx}{dt} = \mu(x) + \eta(t) \; ,$$

with noise as above. The mean $\langle x \rangle$ evolves according to:

$$\frac{d\langle x \rangle}{dt} = \langle \mu(x) \rangle \; .$$

Using an Ito discretization scheme:

$$x(t + \Delta t) = x(t) + \mu(x(t))\Delta t + d\eta(t) \; ,$$

where $d\eta(t) \sim \mathcal{N}(0, D\Delta t)$, we have:

$$x(t + \Delta t) - \langle x(t + \Delta t) \rangle = x(t) - \langle x(t) \rangle + (\mu(x(t)) - \langle \mu(x(t)) \rangle)\Delta t + d\eta(t) \; . \tag{A2}$$

From these, we derive evolution equations for the moments $\langle (x - \langle x \rangle)^n \rangle$ for $n \geq 2$:

$$\frac{d\langle (x - \langle x \rangle)^n \rangle}{dt} = \lim_{\Delta t \to 0} \left\langle \frac{(x(t + \Delta t) - \langle x(t + \Delta t) \rangle)^n}{\Delta t} - \frac{(x(t) - \langle x(t) \rangle)^n}{\Delta t} \right\rangle \; . \tag{A3}$$

Substituting Equation (A2) into the above and simplifying leads to:

$$\frac{d\langle (x-\langle x\rangle)^n\rangle}{dt} = n\langle (x-\langle x\rangle)^{n-1}(\mu(x)-\langle \mu(x)\rangle)\rangle + \binom{n}{2} D\langle (x-\langle x\rangle)^{n-2}\rangle \,. \tag{A4}$$

Now, we re-express:

$$\mu(x) = \mu(x') + \mu'(x')(x-x') + \delta(x,x')(x-x')^2 \,,$$

where δ is at most $O(1)$ in $x-x'$. Then:

$$\begin{aligned}\frac{d\langle (x-\langle x\rangle)^n\rangle}{dt} &= n\mu'(x')\langle (x-\langle x\rangle)^{n-1}(x-\langle x\rangle)\rangle + n\langle (x-\langle x\rangle)^{n-1}(\delta-\langle\delta\rangle)\rangle + \binom{n}{2} D\langle (x-\langle x\rangle)^{n-2}\rangle \\ &= n\mu'(x')\langle (x-\langle x\rangle)^{n}\rangle + n\langle (x-\langle x\rangle)^{n+1}(\delta-\langle\delta\rangle)\rangle + \binom{n}{2} D\langle (x-\langle x\rangle)^{n-2}\rangle \,.\end{aligned}$$

When $\mu'(x') = 0$ and $\delta = 0$ the Green's function is a Gaussian with zero mean and variance Dt, so that $\langle (x-\langle x\rangle)^n\rangle \propto (Dt)^{n/2}$. Inspired by this base case, we consider the moments of the variable $z = (x-\langle x\rangle)/\sqrt{Dt}$:

$$\begin{aligned}\frac{d\langle z^n\rangle}{dt} &= -\frac{n}{2t}\langle z^n\rangle + (Dt)^{-n/2}\frac{d(x-\langle x\rangle)^n}{dt} \\ &= -\frac{n}{2t}\langle z^n\rangle + n\mu'(x')\langle z^n\rangle + n\sqrt{Dt}\langle z^{n+1}(\delta-\langle\delta\rangle)\rangle + \binom{n}{2}\frac{\langle z^{n-2}\rangle}{t} \,.\end{aligned} \tag{A5}$$

We expand $\langle z^n\rangle$ in terms of t, since we are interested in the small-t limit:

$$\langle z^n\rangle = C_n + \alpha_n\sqrt{t} + \beta_n t + \gamma_n t^{3/2} + O(t^2) \,. \tag{A6}$$

In terms of these coefficients, we have:

$$\frac{d\langle z^n\rangle}{dt} = \frac{\alpha_n}{2\sqrt{t}} + \beta_n + \frac{3}{2}\gamma_n\sqrt{t} + O(t) \,. \tag{A7}$$

Substituting Equations (A6) and (A7) into Equation (A5) and matching $O(1/t)$ terms, $O(1/\sqrt{t})$ terms, and so on, yields:

$$0 = -\frac{n}{2}C_n + \binom{n}{2}C_{n-2} \,, \tag{A8}$$

$$\frac{\alpha_n}{2} = -\frac{n}{2}\alpha_n + \binom{n}{2}\alpha_{n-2} \,, \tag{A9}$$

and

$$\beta_n = -\frac{n}{2}\beta_n + n\mu'(x')C_n + \binom{n}{2}\beta_{n-2} \,, \tag{A10}$$

for $O(1/t)$, $O(1/\sqrt{t})$, and $O(1)$, respectively. Note that none of C_n, α_n, or β_n have information about δ, which encapsulates higher-order drift nonlinearities. The $O(\sqrt{t})$ term finally has information about δ:

$$\frac{3}{2}\gamma_n = -\frac{n}{2}\gamma_n + n\mu'(x')\alpha_n + n\sqrt{Dt}\delta(x=x')C_n + \binom{n}{2}\gamma_{n-2}.$$

Interestingly, this implies that any dependencies of the moments on δ are $O(t^{3/2})$, at most. Equations (A8)–(A10) can be solved with the following initial conditions:

$$\langle z^0 \rangle = 1 \rightarrow C_0 = 1,\ \alpha_0 = 0,\ \beta_0 = 0$$

and, by construction:

$$\langle z^1 \rangle = 0 \rightarrow C_1 = 0,\ \alpha_1 = 0,\ \beta_1 = 0\ .$$

Then, $C_n = \alpha_n = \beta_n = 0$ for n odd, and $\alpha_n = 0$ for n even, as well. Some algebra shows that:

$$C_n = \begin{cases} \frac{n!}{(n/2)!2^{n/2}} & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$$

$$\alpha_n = 0$$

$$\beta_n = \begin{cases} \frac{n}{2} C_n \mu'(x') & n \text{ even} \\ 0 & \text{n odd} \end{cases}\ .$$

A Gaussian with mean zero and variance $C_2 + \alpha_2\sqrt{t} + \beta_2 t = 1 + \mu'(x')t$ would also have $C_n = \alpha_n = \beta_n = 0$ for n odd, $\alpha_n = 0$ for n even, and $\langle z^n \rangle_q = C_n(1 + \beta_2 t)^{n/2} = C_n + \frac{n}{2} C_n \mu'(x')t + O(t^2)$. Thus, the moments z^n of $p(z|x')$ are consistent with the moments of $q(z|x')$ to $O(t^{3/2})$. Additionally, as described earlier, those moments are consistent with the moments of the linearized Langevin equation to $o(t)$. From prior logic, $H[p]$ can be approximated to $o(t)$ by $\frac{1}{2}\log(2\pi e|Dt + \mu'(x')Dt^2|)$.

The n-dimensional case follows the same principle, but the calculations are more arduous. We start with the stochastic differential equation for $x \in \mathbb{R}^n$:

$$\frac{dx}{dt} = \mu(x) + \eta(t)\ ,$$

with the noise as before. The initial condition is $x(t = 0) = x'$. Since we are interested not only in whether the distribution is effectively Gaussian, but also in how important the nonlinearities of $\mu(x)$ are, we re-express $\mu(x)$ as:

$$\mu(x) = \mu(x') + A(x')(x - x') + f(x)\ ,$$

where $A_{ij}(x') = \partial\mu_j/\partial x_i$:

$$f_i(x) = \sum_{j,k} \delta_{ijk}(x_j - x'_j)(x_k - x'_k)\ , \tag{A11}$$

and δ_{ijk} is at most of $O(1)$ in $||x - x'||$. The evolution equation for the means is:

$$\frac{d\langle x \rangle}{dt} = \mu(x') + A(x')(\langle x \rangle - x') + \langle f(x) \rangle\ .$$

Using an Ito discretization scheme with time step Δt:

$$x(t + \Delta t) = x(t) + \mu(x')\Delta t + A(x')(x - x')\Delta t + f(x)\Delta t + d\eta(t)\ ,$$

where $d\eta(t) \sim \mathcal{N}(0, D\Delta t)$. From this, we find evolution equations for the moments of x. As before, we subtract the mean:

$$x(t+\Delta t) - \langle x(t+\Delta t)\rangle = x(t) - \langle x(t)\rangle + A(x')(x(t) - \langle x(t)\rangle)\Delta t + (f(x) - \langle f(x)\rangle)\Delta t + d\eta(t) \ . \tag{A12}$$

For notational ease, let $\sigma(1), ..., \sigma(m)$ be a list of integers in the set $\{1, ..., n\}$ where n is the dimension of x; repeats are allowed. We want an evolution equation for $\text{Cov}(x_{\sigma(1)}, ..., x_{\sigma(m)})$:

$$\frac{d}{dt}\text{Cov}(x_{\sigma(1)}, ..., x_{\sigma(m)}) = \frac{d}{dt}\left\langle \prod_{i=1}^{m}(x_{\sigma(i)} - \langle x_{\sigma(i)}\rangle)\right\rangle \ .$$

Using Equation (A12) and steps similar to those outlined in Equations (A3) and (A4), we find that:

$$\begin{aligned}\frac{d}{dt}\text{Cov}(x_{\sigma(1)}, ..., x_{\sigma(m)}) = &\sum_{i=1}^{m}\sum_{k=1}^{n} A_{ik}\text{Cov}(x_{\sigma(k)}, x_{\sigma(j),\ j\neq i}) \\ &+ \sum_{i=1}^{m}\left\langle (f_{\sigma(i)}(x) - \langle f_{\sigma(i)}(x)\rangle)\prod_{j\neq i}(x_{\sigma(j)} - \langle x_{\sigma(j)}\rangle)\right\rangle \\ &+ \sum_{i,j=1}^{m}\text{Cov}(x_{\sigma(k):k\neq i,j}) \ .\end{aligned} \tag{A13}$$

$\text{Cov}(x_{\sigma(k):k\neq i,j})$ denotes the covariance of the variables $x_{\sigma(k)}$ for all k in the integer list $1, ..., m$ with the restriction that we ignore $k = i$ and $k = j$. We have a base case: when $f = 0$, $A = 0$ and $D_{ij} = D\delta_{i,j}$, the Green's function is a Gaussian with variance $\propto \sqrt{t}$. Therefore, again, we switch to variable $z = (x - \langle x\rangle)/\sqrt{t}$ and calculate its covariance evolution, similarly to Equation (A7), where we employ Equation (A11) to find the appropriate t scaling of the nonlinear f term:

$$\begin{aligned}\frac{d\text{Cov}(z_{\sigma(1)}, ..., z_{\sigma(m)})}{dt} = &-\frac{m}{2t}\text{Cov}(z_{\sigma(1)}, ..., z_{\sigma(m)}) + \sum_{i=1}^{m}\sum_{k=1}^{n} A_{ik}\text{Cov}(z_{\sigma(k)}, z_{\sigma(j),\ j\neq i}) \\ &+ \sqrt{t}\sum_{i,j,k}\langle \delta_{ijk} z_{\sigma(j)} z_{\sigma(k)}\prod_{l\neq i} z_{\sigma(l)}\rangle + \frac{1}{t}\sum_{i,j} D_{ij}\text{Cov}(z_{\sigma(k):k\neq i,j}) \ .\end{aligned} \tag{A14}$$

We expand the covariances as a series in $\sqrt{t}$, assuming that they are indeed expressible for short times using such an expansion:

$$\text{Cov}(z_{\sigma(1)}, ..., z_{\sigma(m)}) = \alpha_{\sigma(1),...,\sigma(m)} + \beta_{\sigma(1),...,\sigma(m)}\sqrt{t} + \gamma_{\sigma(1),...,\sigma(m)}t + O(t^{3/2}) \ .$$

As before, we substitute the above series expansion into Equation (A14) and match terms of $O(\frac{1}{t})$, $O(\frac{1}{\sqrt{t}})$, and $O(1)$ to get:

$$\begin{aligned}0 &= -\frac{m}{2}\alpha_{\sigma(1),...,\sigma(m)} + \sum_{i,j} D_{i,j}\alpha_{\sigma(k):k\neq i,j} \ , \\ 0 &= -\frac{m+1}{2}\beta_{\sigma(1),...,\sigma(m)} + \sum_{i,j} D_{i,j}\beta_{\sigma(k):k\neq i,j} \ ,\end{aligned}$$

and

$$0 = -\gamma_{\sigma(1),...,\sigma(m)} - \frac{m}{2}\gamma_{\sigma(1),...,\sigma(m)} + \sum_{i,k} A_{ik}\alpha_{\sigma(k),\sigma(j):j\neq i} + \sum_{i,j} D_{ij}\beta_{\sigma(k):k\neq i,j} \ .$$

The base case is that, by definition, $\langle z \rangle = 0$ and $\langle z^0 \rangle = 1$. This implies that $\beta_{\sigma(1),...,\sigma(m)} = 0$ for all lists $\{\sigma(i) : i = 1, ..., m\}$. Since all moments are determined to at least $O(t)$ by just the linearized version of the nonlinear Langevin equation and since linear Langevin equations have Gaussian Green's functions, it follows that the Green's function for the nonlinear Langevin equation is Gaussian to $O(t)$. Some algebra shows that the variance of the linearized Langevin equation's Green's function is:

$$\text{Var}(q(x|x')) = Dt + \frac{A(x')D + DA(x')^\top}{2} t^2 + O(t^3) \, .$$

If D is invertible, the conditional entropy is then:

$$\begin{aligned} H[X_{t+\tau}|X_t = x'] &= \log \sqrt{2\pi e |\det(D\tau)|} + \frac{1}{2} \log\det\left(I + \frac{D^{-1}A(x')D + A(x')^\top}{2}\tau\right) + O(\tau^2) \\ &= \log \sqrt{2\pi e |\det(D\tau)|} + \frac{1}{2}\text{tr}\left(\frac{D^{-1}A(x')D + A(x')^\top}{2}\tau\right) + O(\tau^2) \\ &= \log \sqrt{2\pi e |\det(D\tau)|} + \frac{\text{tr}(A(x'))}{2}\tau + O(\tau^2) \, . \end{aligned}$$

If the matrix D is not invertible because $\det D = 0$, then we only have the leading order term in t of the entropy $H[p]$ and we cannot draw any conclusions about the $O(1)$ term in any of the information anatomy quantities. This becomes very clear by example in Appendix D.

Now, we return to the question of whether or not we can circumvent the issue of using a moment expansion to approximate entropies of a probability distribution function whose support is not bounded. The key idea is that we are only interested in potential functions $U(x)$ that grow quickly enough with $||x||$, such that the partition function $Z = \int e^{-U(x)} dx$ is normalizable. This suggests that we can approximate the potential function arbitrarily well by a potential function whose support is bounded such that transition probabilities are uniquely determined by moments. Consider, for instance, the sequence of potentials $U_L(x)$ defined by:

$$U^{(L)}(x) = \begin{cases} U(x) & ||x|| \leq L \\ \infty & ||x|| > L \end{cases} \, . \tag{A15}$$

By construction, the transition probabilities have support over the bounded region $||x|| \leq L$. For any of these potentials, the moment expansion above uniquely determines the transition probability distribution. Therefore, the manipulations above give a corresponding sequence of conditional entropies:

$$H^{(L)}[X_{t+\tau}|X_t] = \log \sqrt{2\pi e |\det(D\tau)|} + \int_{||x|| \leq L} \rho_{eq}^{(L)}(x') \frac{\text{tr}(A(x'))}{2} \tau dx' + o(\tau) \, , \tag{A16}$$

where:

$$\rho_{eq}^{(L)}(x) = \begin{cases} \frac{e^{-U(x)}}{\int_{||x'|| \leq L} e^{-U(x')} dx'} & ||x|| \leq L \\ 0 & ||x|| > L \end{cases} \, . \tag{A17}$$

If $\lim_{L\to\infty} H^{(L)}[X_{t+\tau}|X_t] = H[X_{t+\tau}|X_t]$ to $o(\tau)$, then we can claim that the formulae in the main text applies, even when the support of the transition probability distribution function is unbounded. To $o(\tau)$, we see that:

$$\lim_{L\to\infty} H^{(L)}[X_{t+\tau}|X_t] = H[X_{t+\tau}|X_t] + o(\tau) \leftrightarrow \lim_{L\to\infty} \frac{\int_{||x||\leq L} e^{-U(x)}\text{tr}(A(x))dx}{\int_{||x||\leq L} e^{-U(x)}dx} = \frac{\int_{\mathbb{R}^n} e^{-U(x)}\text{tr}(A(x))dx}{\int_{\mathbb{R}^n} e^{-U(x)}dx}. \tag{A18}$$

Thus, we want to know the conditions under which the latter limit converges. In the main text, we limited ourselves to certain types of potential functions, stipulating that $Z = \int_{\mathbb{R}^n} e^{-U(x)}dx < \infty$, so that there is a normalizable equilibrium probability distribution. We also stipulate that $\frac{1}{Z}\int_{\mathbb{R}^n} \text{tr}(A(x))e^{-U(x)}dx < \infty$, so that the bound information rate would be finite. Hence, both $\lim_{L\to\infty}\int_{||x||\leq L} e^{-U(x)}dx = Z < \infty$ and $\lim_{L\to\infty}\int_{||x||\leq L} e^{-U(x)}\text{tr}(A(x))dx = \int_{\mathbb{R}^n} e^{-U(x)}\text{tr}(A(x))dx < \infty$. Since both of these converge to finite, nonzero values, the ratio of the limit is the limit of the ratios, and we have:

$$\lim_{L\to\infty} \frac{\int_{||x||\leq L} e^{-U(x)}\text{tr}(A(x))dx}{\int_{||x||\leq L} e^{-U(x)}dx} = \frac{\int_{\mathbb{R}^n} e^{-U(x)}\text{tr}(A(x))dx}{\int_{\mathbb{R}^n} e^{-U(x)}dx}. \tag{A19}$$

Therefore, this sketch suggests that we can circumvent concerns about using moment expansions. Again, we require that the stationary probability distribution and bound information rate exist and are finite.

D. Linear Langevin Dynamics with Noninvertible Diffusion Matrix

If the stochastic differential equation is linear:

$$\frac{dx}{dt} = A + Bx + \eta(t)\,, \tag{A20}$$

where $\eta(t)$ is white noise $\langle\eta(t)\rangle = 0$ and $\langle\eta(t)\eta(t')^\top\rangle = D\delta(t-t')$, then we can solve it in terms of $\eta(t)$ as follows:

$$\begin{aligned}
\frac{dx}{dt} - Bx &= A + \eta(t) \\
\frac{d}{dt}(e^{-Bt}x) &= e^{-Bt}A + e^{-Bt}\eta(t) \\
e^{-Bt}x(t) - x(0) &= \int_0^t e^{-Bt'}Adt' + \int_0^t e^{-Bt'}\eta(t')dt'\,,
\end{aligned}$$

yielding:

$$x(t) = e^{Bt}x(0) + \int_0^t e^{B(t-t')}Adt' + \int_0^t e^{B(t-t')}\eta(t')dt'\,.$$

Since $\eta(t)$ is white, $x(t)$ is a Gaussian random variable with mean:

$$\langle x(t)\rangle = e^{Bt}x(0) + \int_0^t e^{B(t-t')}Adt'$$

and variance:

$$\begin{aligned}\text{Var}(x(t)) &= \langle (x(t) - \langle x(t) \rangle)(x(t) - \langle x(t) \rangle)^\top \rangle \\ &= \left\langle \int_0^t e^{B(t-t')} \eta(t') dt' \int_0^t \eta(t'')^\top e^{B^\top (t-t'')} dt'' \right\rangle \\ &= \int_0^t e^{B(t-t')} D e^{B^\top (t-t')} dt' \\ &= \int_0^t e^{Bt'} D e^{B^\top t'} dt' \ . \end{aligned} \tag{A21}$$

Since the Green's function is Gaussian for all time (not approximately in the short time limit) and since the variance of this Gaussian does not depend on the initial state, we can calculate the conditional entropies $H[X_t|X_0]$ via:

$$H[X_t|X_0] = \frac{1}{2} \log(2\pi e |\det \text{Var}(x(t))|) \ . \tag{A22}$$

The goal here is to calculate this quantity for small t when the matrix D is not invertible. We assume that it has the block matrix form:

$$D = \begin{pmatrix} 0 & 0 \\ 0 & D_{nn} \end{pmatrix} \ ,$$

where $D_{nn}^\top = D_{nn}$. Let B have the corresponding block matrix form:

$$B = \begin{pmatrix} B_{dd} & B_{dn} \\ B_{nd} & B_{nn} \end{pmatrix} \ .$$

(Recall subscript d stands for deterministic and subscript n for noisy.) We can rewrite the variance in Equation (A21) as a power series in t:

$$\begin{aligned}\text{Var}(x(t)) &= \int_0^t e^{Bt'} D e^{B^\top t'} dt' \\ &= \int_0^t \left(\sum_{k=0}^\infty \frac{B^k}{k!} (t')^k \right) D \left(\sum_{j=0}^\infty \frac{(B^\top)^j}{j!} (t')^j \right) dt' \\ &= \sum_{k,j=0}^\infty \frac{B^k D (B^\top)^j}{k! j!} \int_0^t (t')^{k+j} dt' \\ &= \sum_{k,j=0}^\infty \frac{B^k D (B^\top)^j}{k! j!} \frac{t^{k+j+1}}{k+j+1} \\ &= \sum_{m=1}^\infty \frac{t^m}{m!} \sum_{k=0}^{m-1} \binom{m-1}{k} B^k D (B^\top)^{m-1-k} \ . \end{aligned} \tag{A23}$$

Since we are concerned about the small-t limit, we consider only the first few terms of this power series and, for reasons that will become clear, we write all steps in block-matrix form. The first term, which is of $O(t)$, is the usual:

$$\begin{aligned} Q_1 &= Dt \\ &= \begin{pmatrix} 0 & 0 \\ 0 & D_{nn} \end{pmatrix} t \ . \end{aligned} \tag{A24}$$

The second term, of $O(t^2)$, has the form:

$$\begin{aligned} Q_2 &= \frac{t^2}{2}(BD + DB^\top) \\ &= \frac{t^2}{2}\begin{pmatrix} 0 & B_{dn}D_{nn} \\ D_{nn}B_{dn}^\top & B_{nn}D_{nn} + D_{nn}B_{nn}^\top \end{pmatrix} . \end{aligned} \tag{A25}$$

The third term, of $O(t^3)$, has the form:

$$\begin{aligned} Q_3 &= \frac{t^3}{6}(B^2D + 2BDB^\top + D(B^\top)^2) \\ &= \frac{t^3}{6}\begin{pmatrix} 0 & B_{dd}B_{dn}D_{nn} + B_{dn}B_{nn}D_{nn} \\ (B_{dd}B_{dn}D_{nn})^\top + (B_{dn}B_{nn}D_{nn})^\top & - \end{pmatrix} \\ &\quad + \frac{t^3}{3}\begin{pmatrix} B_{dn}D_{nn}B_{dn}^\top & B_{dn}D_{nn}B_{nn}^\top \\ B_{nn}D_{nn}B_{dn}^\top & - \end{pmatrix} . \end{aligned} \tag{A26}$$

We place a dash in the lower right block matrix entry, since, as it turns out, it does not matter for this calculation. The fourth term, of $O(t^4)$, has the form:

$$\begin{aligned} Q_4 &= \frac{t^4}{24}(B^3D + 3B^2DB^\top + 3BD(B^\top)^2 + D(B^\top)^3) \\ &= \frac{t^3}{6}\begin{pmatrix} (B_{dd}B_{dn} + B_{dn}B_{nn})D_{nn}B_{dn}^\top & - \\ - & - \end{pmatrix} + \frac{t^3}{6}\begin{pmatrix} B_{dn}D_{nn}(B_{dd}B_{dn} + B_{dn}B_{nn})^\top & - \\ - & - \end{pmatrix} . \end{aligned} \tag{A27}$$

Similar to the Q_3 calculation, we care only about the upper left hand entry, and so, every other matrix entry can be ignored. Substituting Equations (A24)–(A27) into Equation (A23), we find that:

$$\text{Var}(x(t)) = \begin{pmatrix} Q_{dd} & Q_{dn} \\ Q_{dn}^\top & Q_{nn} \end{pmatrix} . \tag{A28}$$

where:

$$\begin{aligned} Q_{nn} &= D_{nn}t + \frac{B_{nn}D_{nn} + D_{nn}B_{nn}^\top}{2}t^2 + O(t^3) \\ Q_{dn} &= \frac{B_{dn}D_{nn}}{2}t^2 + \frac{B_{dd}B_{dn}D_{nn} + B_{dn}B_{nn}D_{nn} + 2B_{dn}D_{nn}B_{nn}^\top}{6}t^3 + O(t^4) \\ Q_{dd} &= \frac{B_{dn}D_{nn}B_{dn}^\top}{3}t^3 + \frac{(B_{dd}B_{dn} + B_{dn}B_{nn})D_{nn}B_{dn}^\top}{8}t^4 + \frac{B_{dn}D_{nn}(B_{dd}B_{dn} + B_{dn}B_{nn})^\top}{8}t^4 + O(t^5) . \end{aligned}$$

To find the determinant of the matrix in Equation (A28), we use:

$$\det \text{Var}(x(t)) = \det Q_{nn} \det(Q_{dd} - Q_{dn}Q_{nn}^{-1}Q_{dn}^\top) . \tag{A29}$$

Since $\det D_{nn} \neq 0$, D_{nn} is invertible:

$$\begin{aligned} \det Q_{nn} &= \det(D_{nn}t) \det(I + \frac{D_{nn}^{-1}B_{nn}D_{nn} + B_{nn}^\top}{2}t + O(t^2)) \\ &= \det(D_{nn}t)(1 + \text{tr}(B_{nn})t + O(t^2)) . \end{aligned} \tag{A30}$$

Again, we have used the fact that:

$$\begin{aligned}\text{tr}(D_{nn}^{-1}B_{nn}D_{nn} + B_{nn}^{\top}) &= \text{tr}(B_{nn}D_{nn}D_{nn}^{-1}) + \text{tr}(B_{nn}^{\top}) \\ &= 2\text{tr}(B_{nn}) \, .\end{aligned}$$

Additionally, since D_{nn} is invertible and symmetric, we can also write:

$$\begin{aligned}Q_{nn}^{-1} &= \left(I + \frac{D_{nn}^{-1}B_{nn}D_{nn} + B_{nn}^{\top}}{2}t\right)^{-1} D_{nn}^{-1}t^{-1} + O(t) \\ &= \frac{D_{nn}^{-1}}{t} - \frac{D_{nn}^{-1}B_{nn} + B_{nn}^{\top}D_{nn}^{-1}}{2} + O(t) \, .\end{aligned}$$

Then:

$$\begin{aligned}Q_{dn}Q_{nn}^{-1}Q_{dn}^{\top} = &\left(\frac{B_{dn}D_{nn}}{2}t^2 + \frac{B_{dd}B_{dn}D_{nn} + B_{dn}B_{nn}D_{nn} + 2B_{dn}D_{nn}B_{nn}^{\top}}{6}t^3\right) \\ &\times \left(\frac{D_{nn}^{-1}}{t} - \frac{D_{nn}^{-1}B_{nn} + B_{nn}^{\top}D_{nn}^{-1}}{2}\right) \\ &\times \left(\frac{B_{dn}D_{nn}}{2}t^2 + \frac{B_{dd}B_{dn}D_{nn} + B_{dn}B_{nn}D_{nn} + 2B_{dn}D_{nn}B_{nn}^{\top}}{6}t^3\right)^{\top} + O(t^5) \, .\end{aligned}$$

With some algebra, this becomes:

$$\begin{aligned}Q_{dn}Q_{nn}^{-1}Q_{dn}^{\top} = &\frac{B_{dn}D_{nn}B_{dn}^{\top}}{4}t^3 - \frac{B_{dn}B_{nn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{nn}^{\top}B_{dn}^{\top}}{8}t^4 \\ &+ \frac{B_{dd}B_{dn}D_{nn}B_{dn}^{\top} + 2B_{dn}D_{nn}B_{nn}^{\top}B_{dn}^{\top}}{12}t^4 \\ &+ \frac{B_{dn}D_{nn}B_{dn}^{\top}B_{dd}^{\top} + 2B_{dn}B_{nn}D_{nn}B_{dn}^{\top}}{12}t^4 \\ &+ \frac{B_{dn}B_{nn}D_{nn}B_{dn}^{\top} + (B_{dn}B_{nn}D_{nn}B_{dn}^{\top})^{\top}}{12}t^4 + O(t^5) \\ = &\frac{B_{dn}D_{nn}B_{dn}^{\top}}{4}t^3 + \frac{B_{dn}B_{nn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{nn}^{\top}B_{dn}^{\top}}{8}t^4 \\ &+ \frac{B_{dd}B_{dn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{dn}^{\top}B_{dd}^{\top}}{12}t^4 + O(t^5) \, .\end{aligned}$$

We assume that $B_{dn}D_{nn}B_{dn}^{\top}$ is invertible; *i.e.*, $\det(B_{dn}D_{nn}B_{dn}^{\top}) \neq 0$. Therefore:

$$\begin{aligned}F &= \det(Q_{dd} - Q_{dn}Q_{nn}^{-1}Q_{dn}^{\top}) \\ &= \det\left(\frac{B_{dn}D_{nn}B_{dn}^{\top}}{12}t^3 + (M_{dd} - M_{dn})t^4\right) + O(t^2) \\ &= \det\left(\frac{B_{dn}D_{nn}B_{dn}^{\top}}{12}t^3\right)\left(1 + 12\text{tr}((B_{dn}D_{nn}B_{dn}^{\top})^{-1}(M_{dd} - M_{dn}))t + O(t^2)\right) \, , \end{aligned} \tag{A31}$$

where:

$$M_{dd} = \frac{B_{dd}B_{dn}D_{nn}B_{dn}^{\top} + B_{dn}B_{nn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{dn}^{\top}B_{dd}^{\top} + B_{dn}D_{nn}B_{nn}^{\top}B_{dn}^{\top}}{8}$$

and:

$$M_{dn} = \frac{B_{dn}B_{nn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{nn}^{\top}B_{dn}^{\top}}{8} + \frac{B_{dd}B_{dn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{dn}^{\top}B_{dd}^{\top}}{12}$$

so that:

$$M_{dd} - M_{dn} = \frac{B_{dd}B_{dn}D_{nn}B_{dn}^{\top} + B_{dn}D_{nn}B_{dn}^{\top}B_{dd}^{\top}}{24}.$$

Liberal application of several identities—$\mathrm{tr}(XY) = \mathrm{tr}(YX)$, $\mathrm{tr}(X) = \mathrm{tr}(X^{\top})$ and $\mathrm{tr}(X+Y) = \mathrm{tr}(X) + \mathrm{tr}(Y)$—reveals:

$$F = \det\left(\frac{B_{dn}D_{nn}B_{dn}^{\top}}{12}t^3\right)\left(1 + \mathrm{tr}(B_{dd})t + O(t^2)\right).$$

Substituting Equations (A30) and (A31) into Equation (A29) and substituting that into Equation (A22), we have the conditional entropy:

$$\begin{aligned} H[X_t|X_0] = &\log\sqrt{|\det D_{nn}|} + \log\sqrt{|\det B_{dn}D_{nn}B_{dn}^{\top}|} + \log\sqrt{2\pi e} \\ &+ \frac{3m+n}{2}\log t - m\log\sqrt{12} + \frac{1}{2}\left(\mathrm{tr}(B_{dd}) + \mathrm{tr}(B_{nn})\right)t + O(t^2)\,. \end{aligned}$$

E. Time-Local Predictive Information

Information anatomy measures should have a broad application to monitoring and guiding the behavior of adaptive autonomous agents. Practically, information anatomy gives a suite of semantically distinct kinds of information [6,40] that is substantially richer and structurally more incisive than simple uses of Shannon mutual information that implicitly assume there is only a single kind of (correlational) information. For example, it is reasonable to hypothesize that biological sensory systems are optimized to transmit with high fidelity information that is predictively useful about stimuli or environmental organization. In such a setting, the bound information quantifies how much predictability is lost if one has extracted the full predictable information $\mathbf{E}$ from the past, but chooses to ignore the present $H[X_0]$. Along these lines, the time-local predictive information (TiPi) was recently proposed as a quantity that agents maximize in order to access different behavioral modes when adapting to their environment [12].

(For clarity, we must address a persistently misleading terminology at use here, since it is critical to correctly interpreting the benefits of information-theoretic analyses. The proposed measure is a special case of bound information b_μ. Recall that both b_μ and the excess entropy $\mathbf{E}$ capture the amount of information in the future that is predictable [5,6] and not that which is predictive. The latter is the amount of information that must be stored to optimally predict, and this is given by the statistical complexity C_μ. Therefore, when we use the abbreviation, TiPi, we mean the time-local *predictable* information: information the agent immediately sees as advantageous.)

In fact, [12] does a calculation very similar to the ones above, considering discrete-time stochastic dynamics of the form:

$$x_t = \phi(x_{t-1}) + \eta_t$$

and calculating the TiPi:

$$I^T[X_t; X_{t-1}] \equiv I[X_t; X_{t-1}|X_{t-T} = x_{t-T}]\,, \tag{A32}$$

with fixed $T > 1$. The motivation being that, whatever the history prior to $t - T$, the agent knows the environment state x_{t-T} then. However, from that time forward, the agent, making no further

observations, is ignorant. The stochastic dynamics then models the evolution of that ignorance from the given state to a distribution of states at $t-1$ and then at t, taking into account only the model ϕ the agent has learned or is given. They report that TiPi is the difference between state information and noise entropy:

$$I^T[X_t; X_{t-1}] = \tfrac{1}{2} \ln |\det \Sigma| - \tfrac{1}{2} \ln |\det D| \,, \tag{A33}$$

where:

$$\begin{aligned} D &= \langle \eta \eta^\top \rangle \,, \\ \Sigma &= \sum_{k=1}^{T} L(x_{t-k}) D L(x_{t-k})^\top \,, \\ (L(x))_{ij} &= \frac{\partial \phi_i(x)}{\partial x_j} \,, \end{aligned} \tag{A34}$$

and:

$$L^{(k)}(t-1) = \prod_{m=1}^{k} L(x_{t-m}) \,,$$

with $L^{(0)} = I$.

Since Σ depends on the states between times $t-T$ and $t-1$, the TiPi expression in Equation (A33) also depends on the states between times $t-T$ and $t-1$. The TiPi definition in Equation (A32) does not. Thus, even though the numerical results of [12] are quite interesting, the quantity that the behavioral agents there were maximizing was not the stated conditional mutual information.

To address this concern and explore informational adaptation hypotheses, let us consider alternatives. If desired, for example, one could define an averaged TiPi as:

$$\begin{aligned} I_1^T[X_t; X_{t-1}] &\equiv I[X_t; X_{t-1} | X_{t-T}] \\ &= H[X_t | X_{t-T}] - H[X_t | X_{t-1}, X_{t-T}] \,. \end{aligned}$$

or one could define TiPi to be:

$$I_2^T[X_t; X_{t-1}] \equiv H[X_t | X_{t-T} = x_{t-T}] - H[X_t | X_{t-1} = x_{t-1}] \,,$$

so that it depends on both x_{t-T} and x_{t-1}.

Even with these modifications, Equation (A33) still cannot be a general expression for TiPi, since it depends on measurements at intermediate times that must be marginalized out of the conditional probability distribution with which we are calculating the mutual information.

Moving to discrete time with a small discretization time, let us find expressions for all three:

$$\begin{aligned} I^N[X_t; X_{t-\tau}] &= I[X_t; X_{t-\tau} | X_{t-N\tau} = x_{t-N\tau}] \\ I_1^N[X_t; X_{t-\tau}] &= I[X_t; X_{t-\tau} | X_{t-N\tau}] \\ I_2^N[X_t; X_{t-\tau}] &= H[X_t | X_{t-N\tau} = x_{t-N\tau}] - H[X_t | X_{t-\tau} = x_{t-\tau}] \,. \end{aligned}$$

Suppose that the underlying dynamical system is a nonlinear Langevin equation with invertible diffusion matrix and an analytic potential function U_θ parametrized by θ:

$$\frac{dx}{dt} = -D\nabla U_\theta(x) + \eta(t) \ ,$$

with white noise: $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t')^\top \rangle = D\delta(t-t')$. Following the argument used in Section 3:

$$H[X_t|X_{t-N\tau}] = \log\sqrt{2\pi e(N\tau)^n|\det D|} - \frac{N\tau}{2}\int \nabla\cdot(D\nabla U_\theta(x))P(X_{t-N\tau}=x)dx + O((N\tau)^2) \ ,$$

$$H[X_t|X_{t-N\tau}=x_{t-N\tau}] = \log\sqrt{2\pi e(N\tau)^n|\det D|} - \frac{N\tau}{2}\nabla\cdot(D\nabla U_\theta(x))|_{x=x_{t-N\tau}} + O((N\tau)^2) \ ,$$

$$H[X_t|X_{t-\tau}] = \log\sqrt{2\pi e\tau^n|\det D|} - \frac{\tau}{2}\int \nabla\cdot(D\nabla U_\theta(x))P(X_{t-\tau}=x)dx + O(\tau^2) \ ,$$

and

$$H[X_t|X_{t-\tau}=x_{t-\tau}] = \log\sqrt{2\pi e\tau^n|\det D|} - \frac{\tau}{2}\nabla\cdot(D\nabla U_\theta(x))|_{x_{t-\tau}} + O(\tau^2) \ .$$

These formulae lead to the following expressions for the TiPi alternatives:

$$\begin{aligned} I^N[X_t;X_{t-\tau}] &= n\log\sqrt{N} - \frac{N\tau}{2}\nabla\cdot(D\nabla U_\theta(x))_{x=x_{t-N\tau}} \\ &\quad + \frac{\tau}{2}\int \nabla\cdot(D\nabla U_\theta(x))P(X_{t-\tau}=x|X_{t-N\tau}=x_{t-N\tau})dx + O((N\tau)^2) \ , \qquad \text{(A35)} \\ I_1^N[X_t;X_{t-\tau}] &= n\log\sqrt{N} - \frac{N\tau}{2}\int \nabla\cdot(D\nabla U_\theta(x))P(X_{t-N\tau}=x)dx \\ &\quad + \frac{\tau}{2}\int \nabla\cdot(D\nabla U_\theta(x))P(X_{t-\tau}=x)dx + O((N\tau)^2) \ , \end{aligned}$$

and

$$I_2^N[X_t;X_{t-\tau}] = n\log\sqrt{N} - \frac{N\tau}{2}\nabla\cdot(D\nabla U_\theta(x))|_{x_{t-N\tau}} + \frac{\tau}{2}\nabla\cdot(D\nabla U_\theta(x))|_{x_{t-\tau}} + O((N\tau)^2) \ .$$

Maximizing these with respect to θ has a different effect on the action policy. Maximizing the original TiPi $I^N[X_t;X_{t-\tau}]$ leads the agent to alter the landscape, so that it is driven into unstable regions. Maximizing the averaged TiPi $I_1^N[X_t;X_{t-\tau}]$ leads to a flattening of the potential landscape. Additionally, the effect of maximizing $I_2^N[X_t;X_{t-\tau}]$ is not yet clear.

Not surprisingly, when N is small, we recover the result that maximizing $I^N[X_t;X_{t-\tau}]$ has the same effect on the potential landscape as maximizing the TiPi in [12] when $T=2$. Though the model there is set up for a discrete-time analysis, it is natural to suppose that adaptive agents in an environment move according to a continuous-time dynamic, but receive sensory signals in a discrete-time manner. Equating notation used here and there:

$$\phi(x) = x - D\nabla U(x)\tau$$

gives:

$$L^{(1)} = I - A(x_{t-\tau})\tau \ ,$$

where $A_{ij} = \partial(D\nabla U(x))_i/\partial x_j$. When $N = 2$, substituting this into Equation (A34) yields:

$$\begin{aligned}\Sigma &= D\tau + L^{(1)}D\tau(L^{(1)})^\top \\ &= 2D\tau - (DA(x_{t-\tau})^\top + A(x_{t-\tau})D)\tau^2 + O(\tau^3)\ .\end{aligned}$$

This then gives, upon substitution into Equation (A33):

$$\begin{aligned}\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|D\tau| &= n\log\sqrt{2} - \mathrm{tr}(A(x_{t-\tau}))\tau + O(\tau^2) \\ &= n\log\sqrt{2} - \nabla\cdot(D\nabla U(x))|_{x_{t-\tau}}\tau + O(\tau^2)\ .\end{aligned}$$

The above expression is identical to that in Equation (A35) for all practical purposes, as derivatives of the two with respect to θ are identical up to an unimportant multiplicative constant to subleading order in τ. Therefore, for $T = 2$, many of the qualitative conclusions from numerical simulations are likely to carry over when Equation (A35) is used as the objective function.

Finally, the difference in how these quantities were calculated is interesting to us. For instance, was the series expansion for the coefficients of the moments of the Green's function in Appendix C actually necessary? Could we have used an Ito discretization scheme to write $x_{t+\Delta t}$ in terms of $x_{t-\Delta t}$ and noise terms and use that expression to evaluate b_μ? This is related to the approach taken in [12]. However, the answer obtained using the moment series expansions is a factor of two different than what would have been obtained with such a discretization scheme. Additionally, by keeping track of the order of the approximation errors in Appendix C, we found that these formulae for both bound information and TiPi would only hold for invertible diffusion matrices. As suggested by Appendix D, our estimates for such conditional mutual information change qualitatively when the diffusion matrix is not invertible. That, in turn, may be relevant to environments that are hidden Markov, settings for which the agent's sensorium does not directly report the environmental states.

References

1. Walters, P. *An Introduction to Ergodic Theory*; Graduate Texts in Mathematics, Volume 79.; Springer-Verlag: New York, NY, USA, 1982;
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2006.
3. Crutchfield, J.P.; Young, K. Inferring Statistical Complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108.
4. Shalizi, C.R.; Crutchfield, J.P. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *J. Stat. Phys.* **2001**, *104*, 817–879.
5. Crutchfield, J.P.; Feldman, D.P. Regularities Unseen, Randomness Observed: Levels of Entropy Convergence. *Chaos* **2003**, *13*, 25–54.
6. James, R.G.; Ellison, C.J.; Crutchfield, J.P. Anatomy of a Bit: Information in a Time Series Observation. *Chaos* **2011**, *21*, 037109.
7. Palmer, S.E.; Marre, O.; Berry, M.J., II; Bialek, W. Predictive Information in a Sensory Population **2013**, arXiv:1307.0225.
8. Beer, R.D.; Williams, P.L. Information Processing and Dynamics in Minimally Cognitive Agents. *Cogn. Sci.* **2014**, in press.

9. Tononi, G.; Edelman, G.M.; Sporns, O. Complexity and Coherency: Integrating Information in the Brain. *Trends Cogn. Sci.* **1998**, *2*, 474–484.
10. Strelioff, C.C.; Crutchfield, J.P. Bayesian Structural Inference for Hidden Processes. *Phys. Rev. E* **2014**, *89*, 042119.
11. Sato, Y.; Akiyama, E.; Crutchfield, J.P. Stability and Diversity in Collective Adaptation. *Physica D* **2005**, *210*, 21–57.
12. Martius, G.; Der, R.; Ay, N. Information driven self-organization of complex robotics behaviors. *PLoS One* **2013**, *8*, e63400.
13. Varn, D.P.; Canright, G.S.; Crutchfield, J.P. Discovering Planar Disorder in Close-Packed Structures from X-Ray Diffraction: Beyond the Fault Model. *Phys. Rev. B* **2002**, *66*, 174110–174113.
14. Varn, D.P.; Canright, G.S.; Crutchfield, J.P. ϵ-Machine spectral reconstruction theory: A direct method for inferring planar disorder and structure from X-ray diffraction studies. *Acta. Cryst. Sec. A* **2013**, *69*, 197–206.
15. Crutchfield, J.P.; Young, K. Computation at the Onset of Chaos. In *Entropy, Complexity, and the Physics of Information*; Zurek, W., Ed.; Volume VIII, SFI Studies in the Sciences of Complexity; Addison-Wesley: Reading, MA, USA, 1990; pp. 223–269.
16. Tchernookov, M.; Nemenman, I. Predictive Information in a Nonequilibrium Critical Model. *J. Stat. Phys.* **2013**, *153*, 442–459.
17. Atmanspracher, H.A.; Scheingraber, H. *Information Dynamics*; Plenum: New York, NY, USA, 1991; pp. 45–60.
18. James, R.G.; Burke, K.; Crutchfield, J.P. Chaos Forgets and Remembers: Measuring Information Creation and Storage. *Phys. Lett. A* **2014**, 378, 2124–2127.
19. Lizier, J.; Prokopenko, M.; Zomaya, A. Information modification and particle collisions in distributed computation. *Chaos* **2010**, *20*, 037109.
20. Flecker, B.; Alford, W.; Beggs, J.M.; Williams, P.L.; Beer, R.D. Partial Information Decomposition as a Spatiotemporal Filter. *Chaos* **2011**, *21*, 037104.
21. Moss, F.; McClintock, P.V.E. *Noise in Nonlinear Dynamical Systems*; Cambridge University Press: Cambridge, UK, 1989; Volume 1.
22. Shraiman, B.; Wayne, C.E.; Martin, P.C. Scaling Theory for Noisy Period-Doubling Transitions to Chaos. *Phys. Rev. Lett.* **1981**, *46*, 935.
23. Crutchfield, J.P.; Nauenberg, M.; Rudnick, J. Scaling for External Noise at the Onset of Chaos. *Phys. Rev. Lett.* **1981**, *46*, 933.
24. Girardin, V. On the Different Extensions of the Ergodic Theorem of Information Theory. In *Recent Advances in Applied Probability Theory*; Baeza-Yates, R., Glaz, J., Gzyl, H., Husler, J., Palacios, J.L., Eds.; Springer: New York, NY, USA, 2005; pp. 163–179.
25. Gaspard, P.; Wang, X.J. Noise, Chaos, and (ϵ, τ)-Entropy Per Unit Time. *Phys. Rep.* **1993**, *235*, 291–343.
26. Oksendal, B. *Stochastic Differential Equations: An Introduction with Applications*, 6th ed.; Springer: New York, NY, USA, 2013.
27. Yeung, R.W. *Information Theory and Network Coding*; Springer: New York, NY, USA, 2008.

28. Gaspard, P. Brownian Motion, Dynamical Randomness, and Irreversibility. *New J. Phys.* **2005**, *7*, 77–90.
29. Lecomte, V.; Appert-Rolland, C.; van Wijland, F. Thermodynamic Formalism for Systems with Markov Dynamics. *J. Stat. Phys.* **2007**, *127*, 51–106.
30. Ellison, C.J.; Mahoney, J.R.; Crutchfield, J.P. Prediction, Retrodiction, and the Amount of Information Stored in the Present. *J. Stat. Phys.* **2009**, *136*, 1005–1034.
31. Crutchfield, J.P.; Ellison, C.J.; Mahoney, J.R. Time's Barbed Arrow: Irreversibility, Crypticity, and Stored Information. *Phys. Rev. Lett.* **2009**, *103*, 094101.
32. Crutchfield, J.P.; Feldman, D.P. Statistical Complexity of Simple One-Dimensional Spin Systems. *Phys. Rev. E* **1997**, *55*, R1239–R1243.
33. Poston, T.; Stewart, I. *Catastrophe Theory and Its Applications*; Pitman: London, UK, 1978.
34. Feldman, D.P.; Crutchfield, J.P. Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy. *Phys. Rev. E* **2003**, *67*, 051103.
35. Kittel, C.; Kroemer, H. *Thermal Physics*, 2nd ed.; W. H. Freeman: New York, NY, USA, 1980.
36. Landauer, R. Dissipation and Noise Immunity in Computation, Measurement, and Communication. *J. Stat. Phys.* **1989**, *54*, 1509–1517.
37. Lohr, W. Properties of the Statistical Complexity Functional and Partially Deterministic HMMs. *Entropy* **2009**, *11*, 385–401.
38. Risken, H. *The Fokker-Planck Equation: Methods of Solution and Applications*, 2nd ed.; Springer: Berlin, Germany, 1996.
39. Drozdov, A.N.; Morillo, M. Expansion for the Moments of a Nonlinear Stochastic Model. *Phys. Rev. Lett.* **1996**, *77*, 3280.
40. Crutchfield, J.P.; Ellison, C.J.; Mahoney, J.R.; James, R.G. Synchronization and Control in Intrinsic and Designed Computation: An Information-Theoretic Analysis of Competing Models of Stochastic Computation. *Chaos* **2010**, *20*, 037105.