Information in Dynamics

Neil Gershenfeld

MIT Media Laboratory

20 Ames St, Cambridge, MA 02139 (neilg@media.mit.edu)

Abstract

The historical search for the fundamental meaning of thermodynamic entropy lead to the discovery of the connection between entropy and information about microscopic dynamics, which in turn motivated the development of the theory of information in communications. In this paper I will review how information theory can profitably be applied back to its roots in dynamics in order to characterize the essential properties of a measured time series.

Entropy was introduced in the familiar modern form ($dS = \delta Q/T$) by Clausius in 1854, building on work by Carnot (1824) and Kelvin (∼1850) to understand the nature of heat and irreversability in thermodynamic systems.[1,2] Boltzmann was dedicated to understanding the microscopic meaning of this macroscopic entropy; influenced by Maxwell's kinetic theory of gases (∼1860) he introduced the relationship $H = \int f(\boldsymbol{r}, \boldsymbol{p}, t) \log f(\boldsymbol{r}, \boldsymbol{p}, t) d^3\boldsymbol{r} d^3\boldsymbol{p}$ (where f is the velocity distribution in a gas) in 1872, and then the more general form $S = k \log W$ (where W is the number of available states) around 1877.

In 1871 Maxwell created his demon,[3] which appeared to violate the second law of thermodynamics by intelligent action. Szilard, in 1929, made the significant step of considering a one-molecule gas that could be on either side of a partition; this introducing the idea of a binary bit of information (which side of the partition the molecule is in) and of using entropy to measure information.[4] Although Szilard missed the crucial role of erase in explaining the demon's (mis)behavior (this was first recognized by Landauer[5]), he had laid the foundation for the development of both reversible computation[6] and of information theory.[7] In 1948 Shannon applied entropy to measure the information content in an arbitrary message, independent of its physical origin, and was thereby able to solve significant outstanding communications problems such as the maximum rate at which a message can be sent through a channel. Information theory has since flourished in engineering practice.[8]

The study of ergodic systems has both benefited from, and contributed to, information theory.[9] More recently, Shaw pointed out the connection between information and dissipative dynamics,[10] and Fraser extended this to develop a framework that I will describe here for characterizing the structure in time series produced by non-linear systems.[11] Although a dynamical system's global structure can perform non-trivial computations,[12,13] analyzing the information evolution associated with the much simpler local behavior is sufficient to answer deep questions about the complexity and predictability of a system. In this paper I will will explain how to understand and measure such information. As well as being quite useful in practice, this application of information theory back to its roots in dynamics provides a simple but clear example of the physical meaning of information.[14]

Assume that a physical system is described by a state vector $\vec{x}$ and governing equations $d\vec{x}/dt = \vec{f}(\vec{x})$ (or $\vec{x}_{n+1} = \vec{f}(\vec{x}_n)$). This need not imply that the system is finite-dimensional; the underlying governing equations may be infinite-dimensional partial differential equations which reduce to a finite-dimensional mode expansion due to dissipation. Let $y(\vec{x}(t))$ be a scalar experimentally–accessible quantity that is a function of the state of the system (such as the temperature or velocity at a point in a fluid convection cell, or the concentration of a particular species in a chemical reaction). The goal is to learn as much as possible about the underlying system given only the time series $y(t)$. The necessary connection between the observed time series and its physical origin is provided by state-space reconstruction.[15−18] Construct a new vector out of lagged copies of the observable $\vec{z}_t = (y_t, y_{t-\tau}, \ldots, y_{t-(d-1)\tau})$, where the time lag τ and the dimension d are parameters that will be discussed shortly. If d is large enough, then for almost any choice of the governing equations $\vec{f}$, the observable $y(\vec{x})$, and the time delay τ, the motion of the vector $\vec{z}_t$ will differ from the underlying state vector $\vec{x}_t$ by no more than a smooth invertible change of coordinates (it is an embedding). If the system's dynamics lie on a D-dimensional manifold, then depending on the complexity of its geometry the number of lags d needed for an embedding will be between D and $2D$. Global topological properties (such as the linking of trajectories) as well as local ones (such the rate of divergence of trajectories) are preserved under an embedding, meaning that questions about the underlying system that can be asked in terms of these invariants can be answered by analyzing the embedded data. I will demonstrate this with a few very simple numerical examples; reference [19] describes the application to a range of experimental data sets, and references [20] and [21] are more general review articles that include details about the results to be cited here.

Assume that the time series has been digitized to lie between 1 and N: $y(t) \in \{1, \ldots, N\}$, and estimate the probability of seeing a particular value of the observable as the number of times that it is measured divided by the total number of samples: $p_1(y) = n_y/n_T$. The entropy of this distribution,

$$H_1(N) = -\sum_{y=1}^{N} p_1(y) \log_2 p_1(y) \quad , \tag{1}$$

measures the average number of bits needed to describe one of the samples. The probability of seeing a point in the embedding space $\vec{z}_t = (y_t, \ldots, y_{t-(d-1)\tau})$ is estimated from the joint probability to see this sequence: $p(\vec{z}) = n_{\vec{z}}/n_T = p_d(y_t, \ldots, y_{t-(d-1)\tau})$, and since successive embedded vectors differ by a shift of the coordinates by one place with a new value added at the end, the probability for seeing a particular sequence of embedded vectors will be $p(\vec{z}_t, z_{t-\tau}, \ldots, \vec{z}_{t-(D-1)\tau}) = p_{d+D}(y_t, \ldots, y_{t-(d+D-1)\tau})$. The joint, or block, entropy of such a sequence for a particular choice of the resolution N, embedding dimension d, and delay τ is

$$H_d(\tau, N) = -\sum_{y_t=1}^{N} \cdots \sum_{y_{t-(d-1)\tau}=1}^{N} p_d(\vec{z}_t) \log_2 p_d(\vec{z}_t) \tag{2}$$

This measures the average number of bits needed to describe the sequence. The mutual information between two lagged variables is $2H_1(\tau, N) - H_2(\tau, N)$; this is equal to the average number of bits that one sample provides about the other. This can be generalized to higher dimensions either by $I_d(\tau, N) = dH_1(\tau, N) - H_d(\tau, N)$ or by $R_d(\tau, N) = H_1(\tau, N) + H_{d-1}(\tau, N) - H_d(\tau, N)$. The redundancy R_d is simply related to the mutual information by $R_d = I_d - I_{d-1}$; it is equal to the extra information added by a new sample. The dependence of the redundancy on the parameters d, τ, and N will provide the desired characterization of the time series.

If successive samples are uncorrelated so that the distribution factors ($p_d(y_t, y_{t-\tau}, \ldots, y_{t-(d-1)\tau}) = p_1(y_t)p_1(t_{t-\tau}) \ldots p_1(y_{t-(d-1)\tau})$) then the joint entropy $H_d(\tau, N) = dH_1(N)$ and therefore $R_d(\tau, N) = 0$. On the other hand, if the signal comes from a deterministic system, then once d is large enough to embed the data further samples will not change the distribution: ($p_d(y_t, \ldots, y_{t-(d-1)\tau}) = p_{d-1}(y_t, \ldots, y_{t-(d-2)\tau})$) and so $R_d(\tau, N) = H_1(N)$. The value of d at which the redundancy becomes non-zero is the minimum dimension for which the dynamics can be embedded (if there is one).

The N dependence can be understood in the context of the generalized dimensions of the probability distribution

$$\begin{aligned} D_q &= \lim_{N\to\infty} \frac{1}{q-1} \frac{\log_2 \sum_{\vec{z}_i} p_d(\vec{z}_i)^q}{\log_2 N} \\ \lim_{q\to 1} D_q &= \lim_{N\to\infty} \frac{-\sum_{\vec{z}_i} p_d(\vec{z}_i) \log_2 p_d(\vec{z}_i)}{\log_2 N} \quad . \\ D_1 &= \lim_{N\to\infty} \frac{H_d(\tau, N)}{\log_2 N} \end{aligned} \tag{3}$$

The generalized dimensions will equal the topological dimension for simple objects (1 for a line, 2 for a surface, ...), and will be non-integer for a fractal distribution; the variation with q measures how singular the distribution is. D_1 is called the information dimension, and is the scaling of the entropy with resolution. The smallest integer larger than a measured dimension is the number of local direction available in the system's state-space, and is therefore equal to the number of degrees of freedom necessary to specify the state of the system. If d is below the minimum embedding dimension then the measured dimension will roughly equal the embedding dimension and so the redundancy will vanish; once d is large enough the redundancy will grow with N as $D_1 \log N$.

The τ dependence is related to the source, or Kolmogorov-Sinai, entropy of the signal:

$$h(\tau, N) = \lim_{N\to\infty} \lim_{d\to\infty} H_d(\tau, N) - H_{d-1}(\tau, N) \quad . \tag{4}$$

The source entropy measures the asymptotic rate of increase of information with block size, and is intimately connected with the local dynamics of the system. The eigenvalues $\{\lambda_i\}$ of the local linearization of the dynamics $\partial \vec{f}/\partial \vec{x}$ averaged over the system's trajectory are called the Lyapunov exponents, and if an initial d-ball of trajectories is followed the Lyapunov exponents are equal to the exponential rate of growth (or contraction) of the principal axes. Since the data is being quantized, the divergence associated with positive exponents will reveal information that was not known in the initial conditions. The number of distinguishable states will be proportional to the growth rate of the volume $V(\tau)$, and the information produced by this growth will be the logarithm of the volume: $\log V(\tau) = \log \Pi_i \exp(\lambda_i^+ \tau) = \tau \sum_i \lambda_i^+$ (where the sum runs over the positive exponents). This motivates Pesin's identity:[20] the source entropy for a deterministic system is equal to the sum of the positive Lyapunov exponents, and will be proportional to the delay time τ for small delays: $h(\tau) = \tau h(1) = \tau \sum_i \lambda_i^+$. The source entropy will reach its asymptotic value once d equals the minimum embedding dimension and the resolution is sufficiently fine to produce a generating partition.[11,20] Therefore, if the dynamics are deterministic, then once d and N are large enough the redundancy will fall off with τ proportional to the source entropy $R_d(\tau, N) = H_1(N) - \tau h(1)$. The source entropy can be no larger than the scalar entropy $H_1(N)$, and so if τ can be made so large that successive samples become uncorrelated due to external noise in the system then the redundancy will vanish.

This completes the characterization of a system by the redundancy: (1) for $\tau = 0$ its dependence on N gives the resolution of the observable, (2) as d is increased, if the redundancy becomes non-zero and if its growth in N becomes independent of d then finite-dimensional dynamics have been recovered by embedding, (3) the decay of the redundancy for small τ measures the source entropy, which will be equal to the sum of the positive Lyapunov exponents for deterministic dynamics, and (4) the τ for which the redundancy falls to the noise floor gives the limit of the predictability of the dynamics for that resolution.

I will demonstrate redundancy calculation with four simple examples: (1) uncorrelated uniform random numbers, (2) the quasiperiodic sum of three incommensurate frequencies $x_n = \sin(n) + \sin(\sqrt{2}n) + \sin(\sqrt{3}n)$, (3) the one-dimensional logistic map $x_{n+1} = \lambda x_n(1-x_n)$ with $\lambda = 4$, and (4) the two-dimensional Henon map $x_{n+1} = y_n + 1 - ax_n^2, y_{n+1} = bx_n$ with $a = 1.4$ and $b = 0.3$. Figure 1 shows the complicated time series for these examples, and Figure 2 shows the power spectra; only the quasiperiodic data is produced by a linear system and hence only its spectra provides a meaningful characterization of the system. Figure 3 plots the redundancy as a function of the resolution and the time delay, for embedding dimensions from 1 to 5 using 10^6 point time series from these sources (the redundancy has been normalized by the scalar entropy $H_1(N)$ so that it will be between 0 and 1). For the random data there is no deterministic structure at any resolution, time scale, or embedding dimension, other than the small deviation due to finite sample size. All of the information in the logistic map can be retrieved with just one time delay (because it is a one-dimensional map), and the rapid decay of the entropy indicates a positive Lyapunov exponent. The plot for the Henon map shows that it needs an extra time delay to be embedded, and that the sum of the positive exponents is smaller. For the quasiperiodic data, once the embedding dimension is large enough there is oscillatory structure as τ is increased because the embedded data will occupy a varying number of observable states as τ moves relative to the natural frequencies of the system, and there is no long-term decay in the redundancy because this system has no positive exponents.

The probability distributions p_d that are necessary to calculate the entropy require sorting the data to count how many times particular sequences of samples occur. Although general sorting requires $N \log N$ steps (where N is the number of data points), because the resolution of the data is known in advance it is possible to sort the data with an algorithm that requires only order N steps. First, the successive samples making up an embedded vector are appended to make one long number ($\vec{z} = (12, 3, 14) \rightarrow 120314)$) and these large numbers will be sorted (this is called a lexicographic sort). Because the number of bits in these numbers is known in advance, they can be sorted without any comparisons by entering them into a binary tree based on their bit sequence (this a variation of a bin sort), and the probabilities for successive embedding dimensions may then be read off the appropriate rows of the tree (this is shown in Figure 4). In building this tree pointers are allocated only for occupied nodes so that the storage requirement reflects the actual occupancy of the tree and will usually grow much slower than the number of data points. There is a penalty for this efficiency: the number of samples in different bins may be quite different, meaning that the errors in the probability estimate will vary, which in turn will bias the entropy estimate; algorithm such as k-D trees may be used to have bins with constant occupancy but they sacrifice the ability to incrementally add new samples with a fixed computational cost per sample.[22]

I have briefly reviewed how the early study of dynamics lead to the association between entropy and information, and ultimately to the discovery of information theory, and how information theory (fortified with state-space reconstruction) can be applied back to characterize dynamical systems. Such measurements make no assumption about the linearity of the system, and hence should have an important place next to spectrum analysis as part of the standard repertoire of techniques used for characterizing measured signals.

References

[1] S.G. Brush, *The Kind of Motion we call Heat, Studies in Statistical Mechanics*, Vol. VI (North-Holland, 1976).

[2] R. Balian, *From Microphysics to Macrophysics*, Vol. I (Springer-Verlag, 1991), p. 123.

[3] H.S. Leff and A.F. Rex, *Maxwell's Demon: Entropy, Information, Computing* (Princeton University Press, 1990).

[4] L. Szilard, Z. f. Physik **53**, 840 (1929).

[5] R. Landauer, IBM J. Res. Dev. **5**, 183 (1961).

[6] C.H. Bennett, IBM J. Res. Dev. **17**, 525 (1973).

[7] C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[8] D. Slepian, *Key Papers in the Development of Information Theory* (IEEE Press, 1974).

[9] K. Petersen, *Ergodic Theory* (Cambridge University Press, 1989).

[10] R. Shaw, Z. Naturforsch. **36a**, 80 (1981).

[11] A.M. Fraser, IEEE Trans. Inf. Theory **35** 245 (1989).

[12] C. Moore, Phys. Rev. Lett. **64**, 2354 (1990).

[13] E. Fredkin and T. Toffoli, Int. J. Theor. Phys. **21**, 905 (1982).

[14] R. Landauer, Physics Today **44**, 23 (1991).

[15] F. Takens, *Springer Lecture Notes in Mathematics*, **898**, 366 (1981).

[16] N.H. Packard, J.P. Crutchfield, J.D. Farmer, and R.S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).

[17] D. Ruelle, personal communication.

[18] T. Sauer, J. Yorke and M. Casdagli, *J. Stat. Phys.* **65**, 579 (1991).

[19] M. Palus, in *Predicting the Future and Understanding the Past*, A.S. Weigend and N.A. Gershenfeld eds. (Addison-Wesley, 1993).

[20] J.P. Eckmann and D. Ruelle, Rev. Mod. Phys., **57**, 617 (1985).

[21] N.A. Gershenfeld and A.S. Weigend, preprint, 1992.

[22] F.P. Preparata and M.I. Shamos, *Computational Geometry, An Introduction* (Springer-Verlag, 1985).

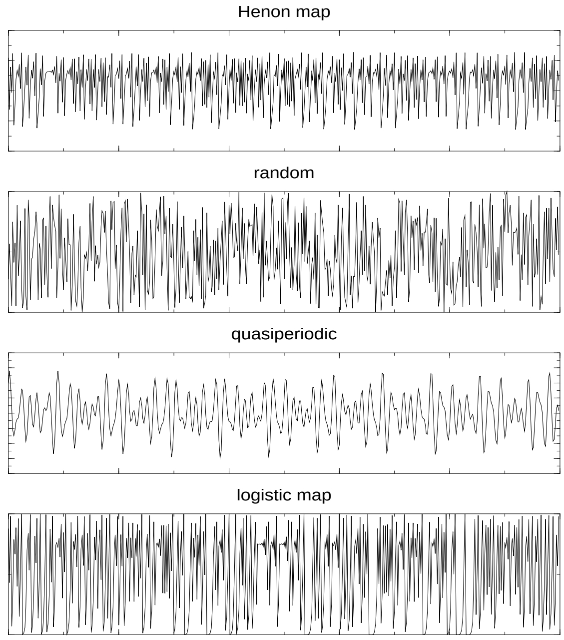

Figure 1: Time series for the four test cases

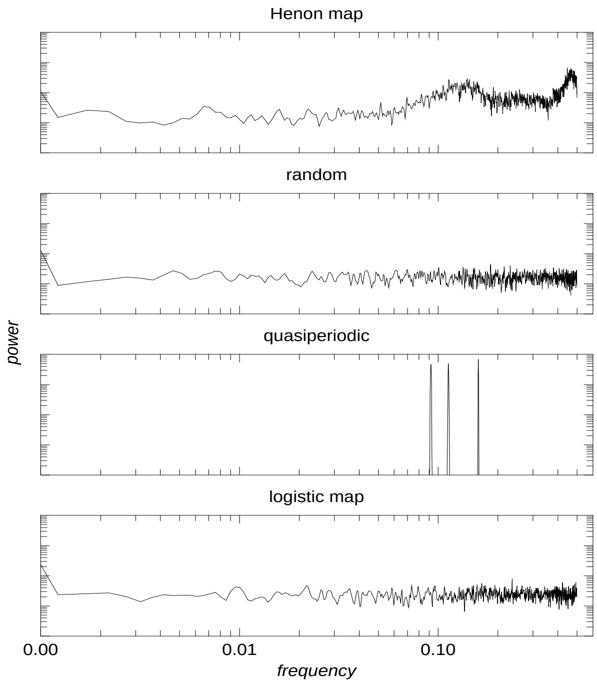

Figure 2: Power spectra for the four test cases

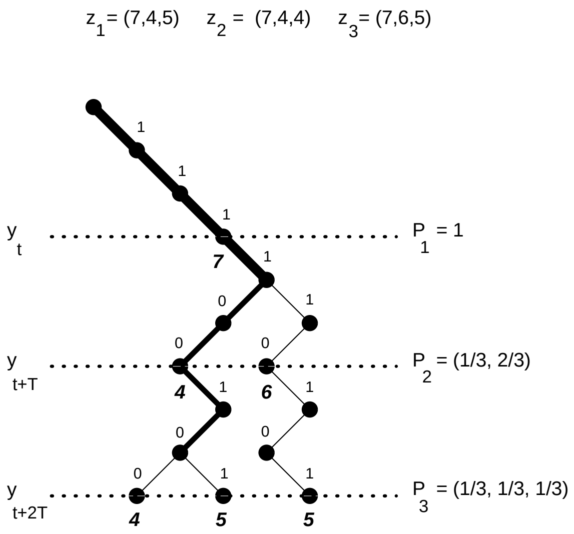

Figure 4: $O(N)$ sorting on a binary tree

Figure 3: Redundancy analyses for the four test cases