# 31 Lecture 31 Information loss rate

### 31.1 Observation with finite resolution

When a dynamical system is given, we wish to describe its state at each instant with a constant precision. Suppose the data with information[357] $H$ is required to describe the current state with a prescribed precision. If we could predict the state with the same precision after one unit time with the aid of this initial data, all the states in any future can be predicted with the same precision in terms of the information $H$ contained in the initial data. However, for chaotic systems the cloud of the ensemble is scrambled and the precise locations of the individual points in the phase space are lost. If we wish to locate the state of the system after one time unit as precisely as we can locate it now, we have to augment $H$ with more information to counter the 'information-dissipating' tendency of chaotic dynamical systems. Thus:
(1) We need extra information $\Delta H$ to describe time one future as accurately as the present. That is, we need more information than $H$ required at present to maintain the precision of the description.
(2) To describe a state at any given time in the steady state we need the same amount $H$ to have 'equi-precision' description. Therefore, $\Delta H$ means the loss of information due to time evolution.

Let us try to quantify the information loss per time.

### 31.2 Interval map information loss rate

Consider a discrete measure-theoretical dynamical system $(F, \mu, I)$ on the interval $I$ with a piecewise continuous and differentiable map[358] $F$ with an absolutely continuous (**29.9**) invariant measure $\mu$. The average amount $h_\mu(F)$ of the extra information required for the equi-precise description of the states is given by the following formula:

$$h_\mu(F) = \int_I \mu(dx)\, \log |F'(x)|. \tag{31.1}$$

Henceforth, we will often write $\mu(dx) = d\mu(x)$.[359] The generalized form that is also

[357] What is information? We only ask how we quantify it. An explanation is given in terms of 'surprisal' in **31.8** and **31.9**.

[358] 'piecewise continuous' implies that a map consists of a several continuous pieces. 'Piecewise continuously differentiable' implies that each piece is differentiable inside, and, at its ends, one-sided derivatives from inside are well-defined.

[359] Here, a formal calculation is explained as given in Y. Oono, "Kolmogorov-Sinai entropy as disorder parameter for chaos," Prog. Theor. Phys. **60**, 1944 (1978). The formula had been obtained

correct for non-absolutely continuous invariant measures is (31.3).[360]

## 31.3 Derivation of Rokhlin's formula

For simplicity, let us assume that $F$ is unimodal (its graph has only one peak and no valley; Fig. 31.1).

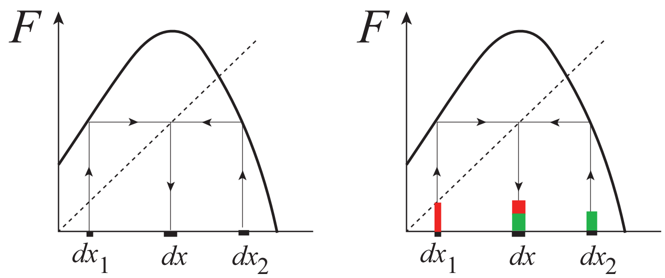


Figure 31.1: An explanation of (31.2). The points in a small interval $dx$ was in $dx_1$ or in $dx_2$ one time step ago. These intervals are stretched and superposed onto $dx$. The right figure illustrates (31.2).

The invariance condition (29.2) for the measure $\mu$ reads, in this case (see Fig. 31.1),

$$\mu(dx_1) + \mu(dx_2) = \mu(dx). \tag{31.2}$$

Here, $dx$ denotes a small interval and its preimage (the set mapped to $dx$ by $F$) is written as $dx_1 \cup dx_2$ ($F^{-1}(dx) = dx_1 \cup dx_2$). For example, if $F$ is applied once to $dx_1$, it is stretched and mapped onto $dx$, so without any doubt the precision to specify the location is reduced. That is, with the equal probability $\mu(dx_1)/\mu(dx) = \mu(dx_1)/\mu(F(dx_1))$ they are spread over $dx$. Consequently, the information was lost by $-\log(\mu(dx_1)/\mu(dx))$. We must also do an analogous calculation for $dx_2$.

Therefore, the expectation value of information deficit required to predict one time step later with the same precision as the current state must be obtained by

---

by Rohlin, "Lectures on the entropy theory of transformations with an invariant measure," Russ. Math. Surveys **22**(5), 1 (1967), so let us call it Rohlin's formula. Note that most important theorems of measure theoretical dynamical systems had been obtained by the Russian mathematicians long before chaos became popular among the US physicists in the 1980s.

[360] F. Ledrappier, "Some properties of absolutely continuous invariant measures on an interval," Ergodic Theor. Dynam. Syst. **1**, 77 (1981) proved the following theorem for $C^2$-maps (actually, for $C^1$-maps with some conditions):

**Theorem** Let $f$ be a $C^2$-endomorphism on an interval. A necessary and sufficient condition for an ergodic invariant measure to be absolutely continuous is that the Kolmogorov-Sinai entropy is given by Rohlin's formula (31.4).

averaging this over $x_1$ with the aid of the (probability) measure $\mu$. Consequently, the rate of information deficit is written as

$$h_\mu(F) = -\int_I \mu(dx) \log \frac{\mu(dx)}{\mu(F(dx))}. \tag{31.3}$$

This loss must be compensated with the extra information ($\Delta H$ in the above) for equi-precise description of the system.

If $\mu$ is absolutely continuous, we can introduce the invariant density $g$ as $g(x)\lambda(dx) = \mu(dx)$, where $\lambda$ is the Lebesgue measure. Hence,

$$\frac{\mu(dx_i)}{\mu(dx)} = \frac{\mu(dx_i)}{\lambda(dx_i)} \frac{\lambda(dx)}{\mu(dx)} \frac{\lambda(dx_i)}{\lambda(dx)} = \frac{g(x_i)}{g(x)|F'(x_i)|}. \tag{31.4}$$

Noting that $F(x_i) = x$ with the aid of (31.4), we may rewrite (31.3) as

$$h_\mu(F) = \int_I \lambda(dx) g(x) \log \frac{g(F(x))|F'(x)|}{g(x)}. \tag{31.5}$$

Here, the integral with respect to the Lebesgue measure is explicitly written as $\int \lambda(dx)$, but it is simply $\int dx$ with the usual notation.

If we apply the Perron-Frobenius equation **31.4**, we see

$$\int_I dy\, g(y) \log g(F(y)) = \int_I dy \int_I dz\, g(y)\delta(z - F(y)) \log g(z) = \int_I dz\, g(z) \log g(z), \tag{31.6}$$

so indeed (31.3) is nothing but (31.1) under the absolute continuity assumption.

## 31.4 Perron-Frobenius equation

If $\mu$ is absolutely continuous, we may introduce the density $g = d\mu/dx \equiv \mu(dx)/\lambda(dx)$. Then, the invariance condition (31.2) for measure $\mu$ reads

$$g(x_1)dx_1 + g(x_2)dx_2 = g(x)dx. \tag{31.7}$$

Here, $F^{-1}(x) = \{x_1, x_2\}$ and $F(dx_i) = dx$ ($i = 1, 2$) (see Fig. 31.1). This means $|F'(x_i)|dx_i = dx$. Therefore, (31.7) reads

$$g(x_1)\frac{dx}{|F'(x_1)|} + g(x_2)\frac{dx}{|F'(x_2)|} = g(x)dx, \tag{31.8}$$

or

$$\frac{g(x_1)}{|F'(x_1)|} + \frac{g(x_2)}{|F'(x_2)|} = g(x). \tag{31.9}$$

With the aid of the elementary property of the $\delta$-function,[361] regarding $y$ as the independent variable, we have for $F$ in **31.3**

$$\delta(x - F(y)) = \frac{1}{|F'(x_1)|}\delta(y - x_1) + \frac{1}{|F'(x_2)|}\delta(y - x_2), \tag{31.10}$$

where $F(x_1) = F(x_2) = x$. Thus can be rewritten as

$$g(x) = \int_I dy\, g(y)\delta(x - F(y)), \tag{31.11}$$

which is called the Perron-Frobenius equation. It is a general formula for the invariant density if a 1D-endomorphism (piecewise differentiable).

### 31.5 Can we solve Perron-Frobenius equation?[362]

Generally, no, but if we assume that the system (defined on [0,1]) allows an absolutely continuous invariant measure, then its density $g$ may be amenable to analytic approach. We start from the Perron-Frobenius equation (31.11). Using Heaviside's step function[363]

$$\Theta(x) = \begin{cases} 0 & x \le 0, \\ 1 & x > 0. \end{cases}, \tag{31.12}$$

the delta function may be written as

$$\frac{d}{dx}\Theta(x - y) = \delta(x - y). \tag{31.13}$$

[361] Let $a$ be an isolated (real) zero point (i.e., $f(a) = 0$), and $f$ be differentiable around it. In a sufficiently small neighborhood of $a$ with the aid of the variable change: $y = f(x)$ we get for sufficiently small positive $\epsilon$

$$\int_{a-\epsilon}^{a+\epsilon} \delta(f(x))\varphi(x)dx = \frac{1}{|f'(a)|}\varphi(a).$$

Here, $\varphi$ is a sufficiently smooth test function. Therefore, if $f$ has a several isolated real zeros $a_i$, we can add each contribution to get

$$\int_{-\infty}^{\infty} \delta(f(x))\varphi(x)dx = \sum_i \frac{1}{|f'(a_i)|}\varphi(a_i).$$

[362] YO talk in Aug 1979 at Res Inst Math Sci, Kyoto

[363] Its definition at $x = 0$ is subtle, but since we do not care for measure zero sets, we may choose it for our convenience.

Introducing this into the Perron-Frobenius equation, we have

$$g(x) = -\int_0^1 dy\, g(y)\frac{1}{F'(y)}\frac{d}{dy}\Theta(x - F(y)). \tag{31.14}$$

Performing an integration by parts, this yields

$$g(x) = \frac{g(0)}{F'(0)}\Theta(x-F(0)) - \frac{g(1)}{F'(1)}\Theta(x-F(1)) + \int_0^1 dy\,\Theta(x-F(y))\frac{d}{dy}\frac{g(y)}{F'(y)}. \tag{31.15}$$

For a map with $F(1) = 0$, we know $g(0) = -g(1)/F'(1)$, this reads

$$g(x) = g(0) + \frac{g(0)}{F'(0)}\Theta(x - F(0)) + \int_0^1 dy\,\Theta(x - F(y))\frac{d}{dy}\frac{g(y)}{F'(y)}. \tag{31.16}$$

Iterative substitution can solve the above equation for unimodal maps,[364] $\beta$-transformations, etc.

$$g(x) = g(0)\left[1 + \sum_{k=1}^{\infty}\frac{\Theta(x - F^k(0))}{[F^k(0)]'}\right]. \tag{31.17}$$

Here the derivative of $F^k(0)$ at the breaking point of $F$ is computed as the left derivative: $[f(x) - f(x-\varepsilon)]/\varepsilon$ $(\varepsilon > 0)$.[365]

### 31.6 KS entropy for billiards

**17.18** explains the general strategy to compute the KS entropy (= information loss rate) and the idea is applied to the Sinai billiard in **17.26**. For these expansive dynamical systems, we have only to pay attention to the stretching rate of the unstable manifolds. We will see this more systematically, when we study Axiom A systems later.

### 31.7 KS entropy of Markov chain

For an ergodic Markov chain[366] the extra information required for equi-precision may

[364] Ito et al. ibid.

[365] to be consistent with our choice of $\Theta$.

[366] See, for example, Durrett, *ibid.*; Z. Brzeźniak and T. Zastawniak, *Basic Stochastic Processes, a course through exercises* (Springer, 1998) is a very kind measure-theoretical introduction to stochastic processes.

be computed by the same idea as explained above. If its transition matrix is given by $\Pi \equiv \{p_{i\to j}\}$, then

$$h(\Pi) = -\sum_{i,j} p_i p_{i\to j} \log p_{i\to j}, \tag{31.18}$$

where $p_i$ is the invariant measure (stationary state) (i.e., $\sum_i p_i p_{i\to j} = p_j$). In particular, for a Bernoulli process $B(p_1, \cdots, p_n)$, we have

$$h(B(p_1, \cdots, p_n)) = -\sum_{i=1}^{n} p_i \log p_i. \tag{31.19}$$

It is easy to understand (31.18). Suppose we are in state $i$ now. After one time step what do we know? With probability $p_{i\to j}$ we will be in state $j$. Thus, we lose on the average the following amount of information

$$-\sum_j p_{i\to j} \log p_{i\to j}. \tag{31.20}$$

We should average this over the probability of our being in state $i$.

## 31.8 Sanov's theorem, a large deviation of observable probabilities

Consider an uncorrelated (= statistically independent) symbol sequence consisting of $n$ symbols. The probability to fine symbol $k$ is given by $q_k > 0$: $\sum q_k = 1$. Suppose we know these probabilities. Observing $N$ symbols, we can obtain the empirical distribution of symbols as

$$\pi_k = \frac{1}{N}\sum_{t=0}^{N-1} \delta_{x_t,k}, \tag{31.21}$$

where $x_t$ is the $t$-th symbol we actually observe.

The law of large numbers for the symbol distribution tells us for any $\varepsilon > 0$

$$\lim_{N\to\infty} P\left(\left\|\frac{1}{N}\sum_{t=0}^{N-1} \delta_{x_t,k} - q_k\right\| > \varepsilon\right) = 0. \tag{31.22}$$

If $N$ is finite, we inevitably observe fluctuations of $\{\pi_k\}$ around $\{q_k\}$. Using the multinomial distribution, we can estimate the probability to observe $\{\pi_k\}$.

If event $i$ occurs $n_i$ times ($\sum_{i=1}^m n_i = N$), the empirical probability is $p_i = n_i/N$.

Therefore, the probability to get this empirical distribution from $N$ sampling is given by the following multinomial distribution:

$$P(\pi \simeq p) = \prod_{i=1}^{m} \frac{N!}{(Np_i)!} q_i^{Np_i}. \tag{31.23}$$

Taking its log and using Stirling's approximation, we obtain

$$\log P(\pi \simeq p) = \log N! + \sum_{i=1}^{m} \{Np_i \log q_i - Np_i \log(Np_i) + Np_i\}. \tag{31.24}$$

That is, we get a large deviation principle for the probability distribution.

$$P(\pi \simeq p) \approx e^{-N\sum_i p_i \log(p_i/q_i)}. \tag{31.25}$$

This relation is called *Sanov's theorem.*[367]

If we introduce the following quantity called the *Kullback-Leibler entropy*

$$K(p\|q) = \sum p_i \log \frac{p_i}{q_i}, \tag{31.26}$$

Sanov's theorem reads

$$P(\pi \simeq p) \approx e^{-NK(p\|q)}. \tag{31.27}$$

As to the nonnegative definite nature of the Kullback-Leibler entropy $K$, see **32.7**.

### 31.9 Information via Sanov's theorem

Since we know the true answer for the symbol distribution, we would not be surprised if $\pi \simeq q$. How should we quantify our surprise?

Suppose we actually observed an event whose probability is $p < 1$. What is the most rational measure of surprise. It is $-\log p$ (or $-\log_2 p$ bits). This is called the surprisal.

> The 'extent of surprise' $f(p)$ we get, spotting a symbol that occurs with probability $p$ or knowing that an event actually happens whose expected probability is $p$, should be

[367] Sanov, I. N. (1957). On the probability of large deviations of random variables, *Mat. Sbornik*, **42**, 11-44. Ivan Nikolaevich Sanov (1919-1968); obituary in (1969). *Russ. Math. Surveys*, **24**, 159.

(1) A monotone decreasing function of $p$ (smaller $p$ should give us bigger surprise).
(2) Nonnegative.
(3) Additive: $f(pq) = f(p) + f(q)$.
Therefore, $f(p) = -c \log p$ $(c > 0)$ is the only choice.[368]

Thus the average surprise per symbol we get is $K(p\|q)$. If we know that the true distribution is flat (even), then the surprisal may be measured by the Shannon formula:

$$H(p) = -\sum_k p_k \log p_k. \tag{31.28}$$

[368] We could invoke the Weber-Fechner law in psychology.