32 Lecture 32. Kolmogorov-Sinai entropy

32.1 Kolmogorov-Sinai entropy as information loss rate

The identity of the information deficiency rate = the extra information required for the equi-precise description and the Kolmogorov-Sinai entropy is generally expected, so the definition of the Kolmogorov-Sinai entropy for a measure-theoretical dynamical system (T, μ, M) is given. This section gives an elementary introduction to the concept and some basic theorems facilitating its intuitive understanding. Some preparation is needed.

32.2 Partition

A (finite) partition $\mathcal{A}$ of a set Γ (Fig. 32.1) is a family of subsets $\{A_1, \cdots, A_n\}$ of Γ satisfying the conditions:

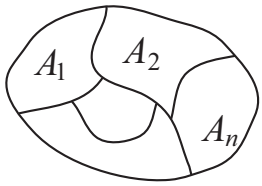

Figure 32.1: Partition $\mathcal{A} = \{A_1, \cdots, A_n\}$. A finite partition $\mathcal{A}$ of a set Γ is a family of finitely many subsets of Γ such that its members A_i do not have any overlap with each other and the total sum perfectly covers Γ.

(1) For any i and j $(\neq i)$ $A_i \cap A_j = \emptyset$, and
(2) $\cup_{i=1}^{n} A_i = \Gamma$

$\{A_i\}$ may be interpreted as the totality of the mutually exclusive observation results.[369] Our observation is always under finite precision. Therefore, if the phase space is continuous, we can never specify a particular point in it by observation. Therefore, it is reasonable to introduce such discrete (coarse-grained) observables.[370]

[369] Here, each set A_i may be a collection of discrete pieces or with holes (i.e., it need not be (singly) connected).

[370] We could not tell which A_i the phase point is actually in, so mustn't A_i be fuzzy? Here, we adopt an interpretation that the relation between the values of a macro-observable that can be observed with a finite precision and the actual microstates of the dynamical system is given by the system itself independent of our capability of observing each microstate. That is, an element of a partition $\mathcal{A}$ consists of a definite set of microstates as an intrinsic property of the system, although we cannot determine this with our finite precision observations.

32.3 Composition of partitions

The composition operation $\vee$ of two partitions of Γ, $\mathcal{A} \equiv \{A_1, \cdots, A_n\}$ and $\mathcal{B} \equiv \{B_1, \cdots, B_m\}$, is defined as follows (Fig. 32.2):

$$\mathcal{A} \vee \mathcal{B} \equiv \{A_1 \cap B_1, A_1 \cap B_2, \cdots, A_n \cap B_m\}. \tag{32.1}$$

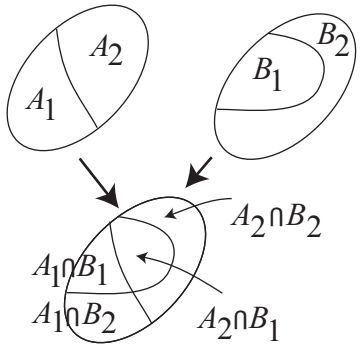

Figure 32.2: Composition of partitions: $\{A_1, A_2\} \vee \{B_1, B_2\} = \{A_1 \cap B_1, A_1 \cap B_2, A_2 \cap B_1, A_2 \cap B_2\}$

On the right-hand side appear all the nonempty pairs of A_i and B_j. By definition $\mathcal{A} \vee \mathcal{B} = \mathcal{B} \vee \mathcal{A}$. The elements of $\mathcal{A} \vee \mathcal{B}$ may be interpreted as the mutually exclusive outcomes obtained by performing two macroscopic observations corresponding to $\mathcal{A}$ and $\mathcal{B}$ simultaneously.

32.4 Information obtainable from observable $\mathcal{A}$

Let us return to the general measure-theoretical dynamical system (T, μ, Γ). The average information we can obtain about the system by a single observation of macroscopic observable $\mathcal{A}$ may be written with the aid of Shannon's formula (**31.9**) as[371]

$$H(\mathcal{A}) = -\sum_{A \in \mathcal{A}} \mu(A) \log \mu(A). \tag{32.2}$$

This information implies that unless we have on the average this amount of information about the system, we cannot infer in which A_i the observation result is in.

[371] The partition must be a measurable partition; $\mu(A_i)$ must be meaningful. Such a statement will not be written explicitly in the following.

32.5 Extra information needed for equi-precise description of observable $\mathcal{A}$

When a system evolves according to the dynamical law T, how much information do we need to infer its coarse-grained state (one of $\mathcal{A}$) at the next time (time $t+1$) under the condition that we know the coarse-grained state at present (time t)? To infer both the states at time t and time $t+1$ without any prior knowledge, we need information of $H(\mathcal{A} \vee T^{-1}\mathcal{A})$, where $T^{-1}\mathcal{A} = \{T^{-1}A_1, \cdots, T^{-1}A_n\}$. We can understand this as follows. An element of $T^{-1}\mathcal{A}$ coincides with some element of $\mathcal{A}$ after one time step. That is, to describe the state of the system after one time step in terms of the observable $\mathcal{A}$ is equivalent to knowing in which element of $T^{-1}\mathcal{A}$ the current state is. Therefore, to specify both the states at $t+1$ and at t is to specify a single element in $\mathcal{A} \vee T^{-1}\mathcal{A}$. Consequently, on the average without $H(\mathcal{A} \vee T^{-1}\mathcal{A})$ of information, we cannot infer in which A_i at time t and in which A_j at $t+1$ the system is in. If we already know the macrostate at time t, to predict the macrostate at $t+1$ we need $H(\mathcal{A} \vee T^{-1}\mathcal{A}) - H(\mathcal{A})$ extra information. This is the amount of extra information ΔH (appearing in **31.1**) required for equi-precision prediction of the one time step future.

32.6 Steady-state information loss

We need not stick to a particular time t and $t+1$ in the steady state. That is, to describe the extent of chaos for a measure-theoretical dynamical system we should consider the average deficiency for a long time:

$$\frac{1}{n}[H(\mathcal{A} \vee T^{-1}\mathcal{A} \vee \cdots \vee T^{-n}\mathcal{A}) - H(\mathcal{A})] \to h_\mu(T, \mathcal{A}), \tag{32.3}$$

where the existence of this limit in $n \to \infty$ is guaranteed by the (intuitively plausible) subadditivity of H **32.7** and $H(T^{-n}\mathcal{A}) = H(\mathcal{A})$ ($n = 1, 2, \cdots$; this can be seen from the invariance of the measure (29.2)).

32.7 Subadditivity of entropy

The following inequality shows the subadditivity of information:

$$H(\mathcal{A} \vee \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B}). \tag{32.4}$$

The inequality must be intuitively natural, because in order to describe two observables $\mathcal{A}$ and $\mathcal{B}$ it is better to use information about the relation between these two as well than to use information separately from each of the observables. Algebraically,

we proceed as follows:

$$H(\mathcal{A}\vee\mathcal{B}) = -\sum_{i,j}\mu(A_i\cap B_j)\log\mu(A_i\cap B_j), \tag{32.5}$$

$$= -\sum_{i,j}\mu(A_i\cap B_j)\left\{\log\left(\frac{\mu(A_i\cap B_j)}{\mu(A_i)\mu(B_j)}\right)+\log\mu(A_i)\mu(B_j)\right\}, \tag{32.6}$$

$$= -\sum_{i,j}\mu(A_i\cap B_j)\log\left(\frac{\mu(A_i\cap B_j)}{\mu(A_i)\mu(B_j)}\right)+H(\mathcal{A})+H(\mathcal{B}). \tag{32.7}$$

If the first term on the right-hand side of (32.7) is non-positive, the proof of $H(\mathcal{A}\vee\mathcal{B})\le H(\mathcal{A})+H(\mathcal{B})$ is over. This is the following important inequality: Let p and q be probabilities ($p_i\ge 0$ and $\sum_i p_i=1$, etc., hold)

$$\sum_i p_i\log\frac{p_i}{q_i}\ge 0. \tag{32.8}$$

To show this we use the inequality $x\log x\ge x-1$ for $x\ge 0$.[372] Introduce $x=p_i/q_i$ into this and sum over i after multiplying q_i:

$$\sum_i q_i\left(\frac{p_i}{q_i}\log\frac{p_i}{q_i}\right)\ge\sum_i p_i-\sum_i q_i=0. \tag{32.9}$$

32.8 Fekete's lemma

If $\{f(n)\}$ is subadditive (i.e., $f(n+m)\le f(m)+f(n)$ holds for any positive integers, n,m), $\lim_{n\to\infty}f(n)/n=\inf_m f(m)/m$.[373]

[Demo] Obviously, $\liminf f(n)/n\ge\inf f(m)/m$. Writing $n=s+km$ ($m>0$, $s\ge 0$ are integers), we get

$$\frac{f(n)}{n}=\frac{f(s+km)}{s+km}\le\frac{f(s)+kf(m)}{s+km}\to\frac{f(m)}{m}. \tag{32.10}$$

[372] The minimum value of $f(x)=x\log x-x+1$ for $x\ge 0$ is zero as can be seen from the graph.

[373] ⟪**Infimum limit (lim inf), supremum limit (lim sup)**⟫ $\liminf_{n\to\infty}x_n=\lim_{n\to\infty}\inf\{x_n,x_{n+1},\cdots\}$. That is, we make the lower bound y_n of the sequence beyond x_n and then take its limit $n\to\infty$. Since $\{y_n\}$ is monotone increasing, the limit is well-defined (may not be bounded); similarly, $\limsup_{n\to\infty}x_n=\lim_{n\to\infty}\sup\{x_n,x_{n+1},\cdots\}$.

Therefore, $\limsup f(n)/n \leq \inf f(m)/m$. Hence, the infimum and supremum limits agree, and the limit exists.

32.9 The best observable: definition of Kolmogorov-Sinai entropy

We should look for the 'best' observable to observe the system. Here, 'the best observable' should imply the observable that allows us to observe the system maximally in detail. Such an observable must be sensitive to the time evolution of the system, so the increasing rate of the information deficiency for this observable must be the largest. With this idea the Kolmogorov-Sinai entropy (or measure-theoretical entropy) is defined as follows:

$$h_\mu(T) \equiv \sup_{\mathcal{A}} h_\mu(T, \mathcal{A}), \tag{32.11}$$

where the supremum is taken over all the finite partitions of Γ (roughly speaking, we try all the finite-resolving power observations).

32.10 Generator

If all the future data of a certain macro-observable determines the future history (trajectory) uniquely, we do not need any more detailed observations. Take an arbitrary history ω. Let us write the element of $\mathcal{A} \vee T^{-1}\mathcal{A} \vee \cdots \vee T^{-n+1}\mathcal{A}$ containing ω as $A^n(\omega)$ (which is an example of a cylinder set; see **26.5**).[374] Since $\omega \in A^n(\omega)$ for any $n = 0, 1, 2, \cdots$, $\omega \in \cap_{n=0}^{\infty} A^n(\omega)$. If this common set does not contain any history other than ω, in other words, if $\cap_{n=0}^{\infty} A^n(\omega) = \{\omega\}$, any further detailed observation is superfluous. If such a relation holds for μ-almost all ω (i.e., except for μ-measure zero set), the partition $\mathcal{A}$ is called a generator. If $\mathcal{A}$ is a generator, as expected,[375]

$$h_\mu(T) = h_\mu(T, \mathcal{A}). \tag{32.12}$$

For example, for $Tx = \{2x\}$ in Appendix 2.1A $\{[0, 1/2], (1/2, 1]\}$ is a generator.

If a measurable partition $\mathcal{A}$ separates all points (precisely speaking, for μ-almost

[374] An element of $\mathcal{A} \vee T^{-1}\mathcal{A} \vee \cdots \vee T^{-n+1}\mathcal{A}$ has the form: $A_i \cap T^{-1}A_j \cap \cdots \cap T^{-n+1}A_k$, which is the totality $\{x : x \in A_i, Tx \in A_j, \cdots, T^{n-1}x \in A_k\}$. Therefore, around here in the text, $x \in \Gamma$ and a history $\omega = \{x, Tx, \cdots\}$ are identified.

[375] P. Walters, *An Introduction to Ergodic Theory* (Springer, 1982) is an excellent textbook of the Kolmogorov-Sinai entropy (but students outside mathematics may not be able to read it with ease). P. Billingsley, *Ergodic Theory and Information* (Wiley, 1960) is also excellent, but is slightly dated.

all points), in other words, for any two points $x \neq y \in \Gamma$, there is a positive integer n and an element $A \in \mathcal{A}$ such that $T^n(x) \in A$ and $T^n(y) \notin A$, $\mathcal{A}$ is a generator. Actually, if $\mathcal{A}$ were not a generator, there are two histories $T^n(x)$ and $T^n(y)$ that are always contained in a single element of $\mathcal{A} \vee T^{-1}\mathcal{A} \vee \cdots \vee T^{-n+1}\mathcal{A}$ for any $n = 0, 1, 2, \cdots$. This contradicts the requirement that each point is separated.

32.11 Chaos and information loss/gain rate

A chaotic dynamical system is a dynamical system for which the extra information required for equi-precise description increases linearly in time.[376]

For chaos, if we wish to predict its state after N time steps with sufficient precision, we need a tremendous amount of information at present (we need $\sim e^{Nh}$ times as much precision as required to describe the present state, so N times as much information as required to know the state at present; see just below), so there is no wonder that sooner or later the behavior of the system becomes unpredictable.

32.12 Krieger's theorem on generator[377]

Let (T, μ, M) be an ergodic dynamical system. If its Kolmogorov-Sinai entropy satisfies $h_\mu(T) < \log k$ for some integer $k > 1$, there is a generator with k elements.

The theorem roughly tells us that if the Kolmogorov-Sinai entropy of the system is $\log k$, then we can encode its dynamics using k symbols without losing any information.

[376] In contrast to the above explanation, there were people who wished to introduce the Kolmogorov-Sinai entropy as the rate of generation of information by the dynamical system. Chaotic dynamical systems often exhibit us details of initial conditions later because of exponential separation of nearby trajectories. Consequently, (if noise is completely ignored) continuous observation of trajectories would tell us increasingly detailed information about the initial condition (in retrospect). In this sense, the dynamical system looks as if it is generating information. This observation itself is correct, but whether this generating rate can be measured by the Kolmogorov-Sinai entropy is another question. For chaos that may be observed numerically (observable chaos), this is correct, but, for example, for unimodal endomorphisms of intervals, except for at most one invariant measure, the assertion does not hold for any of (uncountably many) invariant measures; generally speaking, the generating rate of information just considered is larger than the Kolmogorov-Sinai entropy. This fact is captured by Ruelle's inequality (33.10) we will encounter later. Therefore, it is quite dangerous to interpret the Kolmogorov-Sinai entropy as the information generating rate.

[377] W. Krieger, "On entropy and generators of measure-preserving transformations," Trans. Amer. Math. Soc. **149**, 453 (1970); corrections *ibid.* **168**, 519 (1972).

32.13 Shannon-McMillan-Breiman's theorem[378]

Let (T, μ, Γ) be an ergodic dynamical system and $\mathcal{A}$ a finite partition of Γ. Let $A^n(x)$ be an element of $\mathcal{A} \vee T^{-1}\mathcal{A} \vee \cdots \vee T^{-n+1}\mathcal{A}$ (a cylinder set of length n) containing $x \in \Gamma$. For μ-almost all x

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log \mu(A^n(x)) \right] = h_\mu(T, \mathcal{A}). \tag{32.13}$$

In the above formula $A^n(x)$ is, as before, the bundle of histories (i.e., cylinder set) that cannot be distinguished from the history with the initial condition x for n time steps with a coarse-grained observation corresponding to the partition $\mathcal{A}$. $\mu(A^n(x))$ is the volume of the initial conditions for the history satisfying the condition (we could interpret it as the volume of the cylinder set). If the observation time span n is increased, the condition becomes increasingly stringent, so the volume decreases exponentially. (32.13) measures how fast the volume of the cylinder set decreases. The more chaotic the system is, the harder trajectories to be close to a particular one starting from x, so this value must become larger.

It may be more intuitive to rewrite (32.13) as

$$\mu(A^n(x)) \sim e^{-nh_\mu(T,\mathcal{A})}. \tag{32.14}$$

32.14 Asymptotic equipartition theorem

A special case of the Shannon-McMillan-Breiman theorem is the asymptotic equipartition (AEP) theorem of information theory:[379]

Let $\{X_i\}$ be a sequence of independently and identically distributed stochastic variables. Let us write the probability for a (consecutive) n samples, $X_1, X_2, \cdots, X_n$, as $p(X_1, \cdots, X_n)$. Then,

$$-\frac{1}{n} \log p(X_1, \cdots, X_n) \to H(X), \tag{32.15}$$

where $H(X)$ is the entropy of the individual stochastic variables. Notice that this is nothing but the weak law of large numbers (footnote 28 in Section 1.2) for the

[378] For a proof of the Shannon-McMillan-Breiman theorem, see, for example, W. Parry, *Entropy and Generators in Ergodic Theory* (Benjamin, New York 1969).

[379] It is quite important in information theory. See T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, 1991) p50-

logarithm of probability. $X_1, X_2, \cdots, X_n, \cdots$ may be interpreted as a history, so the Shannon-McMillan-Breiman theorem is the extension of the asymptotic equipartition theorem to histories with correlated events.

32.15 Brin-Katok's theorem

The Brin-Katok theorem explicitly counts the number of histories in the ε-neighborhood of the trajectory starting from x.[380]

Let the totality of the initial conditions that stay in the ε-neighborhood of the trajectory starting from $x \in \Gamma$ for N time steps be (the theorem holds for continuous dynamical systems as well)

$$B_N(x, \varepsilon) = \{y \in M : d(T^n x, T^n y) \leq \varepsilon, 0 \leq n \leq N\}. \tag{32.16}$$

Theorem [Brin-Katok] Let (T, μ, Γ) be an ergodic dynamical system. For μ-almost all $x \in \Gamma$,

$$h_\mu(T) = -\lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{1}{N} \log \mu(B_N(x, \varepsilon)). \tag{32.17}$$

32.16 (some) Chaos can be predictable

Chaos may be predictable for a certain time span thanks to its deterministic nature, in contradistinction to noise, but not for a long time. We can estimate during how many steps n we can predict dynamics with the aid of the Kolmogorov-Sinai entropy h. According to the Brin-Katok theorem $n \sim (\log \delta x)/h$, where δx is our resolving power of the initial condition. However, there are chaotic dynamical systems with very small h, for which accurate prediction of considerable future is possible.[381]

32.17 Chaos under noise

As was stated around the Shannon-McMillan-Breiman theorem above, if a system is chaotic, the bundle of histories close to a given history thins quickly, so it is often the case that the external noise makes a chaotic system 'more chaotic.' This is

[380] M. Brin and A. Katok, "On local entropy" Lecture Notes Math. **1007**, 30 (1983).

[381] For asteroid 522 Helga the Lyapunov characteristic time (the time needed to magnify the initial error by e) is 6,900 years, so accurate prediction is possible. See A. Milani and A. M. Nobili, "An example of stable chaos in the Solar System," Nature, **357**, 569 (1992). J. J. Lissauer, "Chaotic motion in the Solar System," Rev. Mod. Phys. **71**, 835 (1999) is a review.

because noise can induce jumps between the elements of a generator that does not happen with a single time step by the intrinsic dynamics. However, with our crude description of dynamical systems so far given, we cannot claim anything general as to the noise response of a chaotic system because the 'effectiveness' of noise strongly depends on its details such as which trajectories actually come close in the phase space. We cannot conclude that a system with a larger Kolmogorov-Sinai entropy is more sensitive to noise. For example, due to noise transition into a particular element $A_i \in \mathcal{A}$ can become disproportionately frequent. If this element is embedded by the intrinsic dynamics into a (small portion) of another element, the Kolmogorov-Sinai entropy could become smaller due to noise. This is indeed the essence of the noise-induced order discovered by Tsuda.[382] More extremely, we could imagine a set D outside the range of the intrinsic steady dynamics B such that the trajectories perturbed into D from B by noise are sent back into a very small subset of B. Thus, we see that the noise effect is very sensitively dependent on the details of the system (Fig. 32.3).

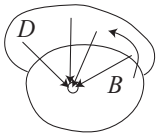

Figure 32.3: A dynamical system that can order by noise. If a phase point goes out of B to D, it is sent back to a very small portion of B. In such a case moderate noise could suppress chaos that is supported on B.

32.18 Open cover

A collection of open sets $\mathcal{O} = \{O_\lambda \subset M, \lambda \in \Lambda\}$ is called an open cover of M, if $M = \cup_{\lambda\in\Lambda} O_\lambda$.

We define the join $\mathcal{O} \vee \mathcal{P}$ of two open covers $\mathcal{O} = \{A_i\}$ and $\mathcal{P} = \{B_j\}$ just as the case of the partitions as (removing empty sets)

$$\mathcal{O} \vee \mathcal{P} = \{A_i \cap B_j\}. \tag{32.18}$$

32.19 Topological entropy

Consider a continuous dynamical system (T, M). Then , if $\mathcal{O}$ is an open cover of M,

[382] K. Matsumoto and I. Tsuda, "Noise-induced order," J. Stat. Mech. **31**, 87 (1983).

$T^{-1}\mathcal{O} \equiv \{T^{-1}A_i\}$ is again an ope cover. Therefore, we can define $\mathcal{O} \vee T^{-1}\mathcal{O}$.

The following limit is called the topological entropy of the open cover $\mathcal{O}$ of (T, M):

$$h_{\text{top}}(T, \mathcal{O}) = \lim_{n\to\infty} \frac{1}{n} \log(\mathcal{O} \vee T^{-1}\mathcal{O} \vee \cdots \vee T^{-n+1}\mathcal{O})^\circ, \tag{32.19}$$

where A° denotes the number of elements in the collection A.

Note that the limit exists just as in the case of te KS entropy, because $(A \vee B)^\circ \leq A^\circ B^\circ$.

The topological entropy of (T, M) is defined as

$$h_{\text{top}}(T) = \sup_{\mathcal{O}} h_{\text{top}}(T, \mathcal{O}). \tag{32.20}$$

Just as in the case of te KS entropy, honestly computing the supremum is hard; there is a counterpart of generators, if M is metrizable.

32.20 Top ent is the lowest upper bound of KS entropies
For a dynamical system (T, M), its topological entropy is given by

$$h_{\text{top}}(T) = \sup_{\mu} h_\mu(T), \tag{32.21}$$

where μ is an invariant measure of T.

32.21 Number of fixed points and topological entropy If (T, M) is topological mixing, and has the pseudo orbit tracing property, then

$$h_{\text{top}}(T) = \limsup_{n\to\infty} \frac{1}{n} \log N(T,\ \text{fixed points}). \tag{32.22}$$

This implies that the convergence radius of the ζ-function is given by $e^{-h_{\text{top}}(T)}$.

For a endomorphism of an interval $F : I \to I$, if topologically mixing, we have only to count the number of peaks of F^n:

$$h_{\text{top}}(F) = \limsup_{n\to\infty} \frac{1}{n} \log N(F,\ \text{fixed points}). \tag{32.23}$$