Teoria Ergódica Diferenciável

lecture 21: Entropy

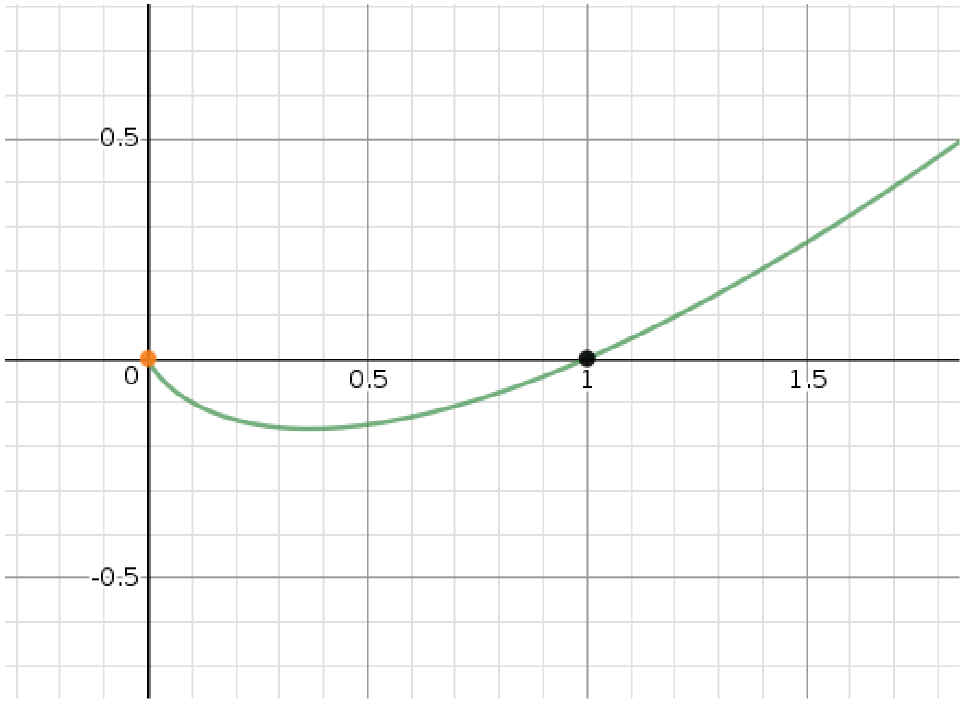
Instituto Nacional de Matemática Pura e Aplicada

Misha Verbitsky, November 29, 2017

Measure-theoretic entropy

DEFINITION: Partition of a probability space (M, μ) is a countable decomposition $M = \coprod V_i$ onto a disjoint union of measurable set. **Refinement** of a partition $\mathcal{V} = \{V_i\}$ is a partition $\mathcal{W}$, obtained by partition of some of V_i into subpartitions. In this case we write $\mathcal{V} \prec \mathcal{W}$. **Minimal common refinement** of partitions $\mathcal{V} = \{V_i\}$, $\mathcal{W} = \{W_j\}$ is a partition $\mathcal{V} \vee \mathcal{W} = \{V_i \cap W_j\}$.

DEFINITION: Entropy of a partition $\mathcal{V} = \{V_i\}$ is $H_\mu(\mathcal{V}) := -\sum_i \mu(V_i) \log(\mu(V_i))$.



EXERCISE: The entropy of infinite partition can be infinite. **Find a partition with infinite entropy.**

Entropy of a communication channel

Consider a communication channel which sends words, chosen randomly of k letters which appear with probabilities $p_1, ..., p_k$, with $\sum_i p_k = 1$. The entropy of this channel is $H(p_1, ..., p_k)$ **measures "informational density" of communication** (C. Shannon).

It should satisfy the following natural conditions.

1. Let $l > k$. The information density is clearly higher for $p_1 = ... = p_k = 1/k$ than for $q_1, ..., q_l = 1/l$. **Therefore, $H(1/k, ..., 1/k) < H(1/l, ..., 1/l)$.**

2. H should be **continuous as a function of p_i** and symmetric under their permutations.

3. Suppose that we have replaced the first letter in the alphabeth of k letters by l letters, appearing with probabilities $q_1, ..., q_l$. We have obtained a communication channel with $k + l - 1$ letters, with probabilities $p_1q_1, ..., p_1q_l, p_2, ..., p_k$. **Then $H(p_1q_1, ..., p_1q_l, p_2, ..., p_k) = H(p_1, ..., p_k) + p_1H(q_1, ..., q_l)$.**

Clearly, $H(p_1, ..., p_k) = -\sum p_i \log p_i$ satisfies these axioms. Indeed,

$$-\sum_{i=2}^{k} p_i \log p_i - \sum_{j=1}^{l} p_1 q_j \log(p_1 q_j) = -\sum_{i=2}^{k} p_i \log p_i - p_1 \log p_1 - p_1 \sum_{j=1}^{l} q_j \log q_j.$$

It is possible to show that $H(p_1, ..., p_k) = -\sum p_i \log p_i$ **is the only function which satisfies these axioms.**

C. Shannon, "Mathematical theory of computation", p. 10

6. Choice, Uncertainty and Entropy

We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is reasonable to require of it the following properties:

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three

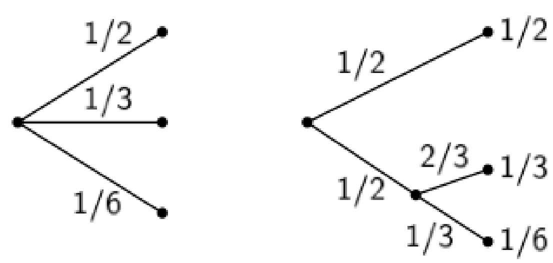


Fig. 6 — Decomposition of a choice from three possibilities.

possibilities $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}, \frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} H(\tfrac{2}{3}, \tfrac{1}{3}).$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

Entropy of dynamical system

In this lecture, we consider only dynamical systems (M, μ, T) with μ probabilistic and T measure-preserving.

Given a partition $\mathcal{V}, M = \coprod V_i$ **we denote by $T^{-1}(\mathcal{V})$ the partition $M = \coprod T^{-1}(V_i)$.**

DEFINITION: Let (M, μ, T) be a dynamical system, and $\mathcal{V}, M = \coprod V_i$ a partition of M. Denote by $\mathcal{V}^n$ the partition $\mathcal{V}^n := \mathcal{V} \vee T^{-1}(\mathcal{V}) \vee T^{-2}(\mathcal{V}) \vee ... \vee T^{-n+1}$. **Entropy (M, μ, T) of with respect to the partition $\mathcal{V}$** is $h_\mu(T, \mathcal{V}) := \overline{\lim}_n \frac{1}{n} H_\mu(\mathcal{V}^n)$ **Entropy of (M, μ, T)** is supremum of $h_\mu(T, \mathcal{V})$ taken over all partitions $\mathcal{V}$ with finite entropy.

REMARK: Let $\mathcal{V} \succ \mathcal{W}$ be a refinement of the partition $\mathcal{W}$. Clearly, $H_\mu(\mathcal{V}) \geqslant H_\mu(\mathcal{W})$. This implies $h_\mu(T, \mathcal{V}) \geqslant h_\mu(T, \mathcal{W})$.

Entropy of dynamical system and iterations

REMARK: Clearly, $\bigvee_{j=0}^{n-1} T^{-j}(\mathcal{V}^k) = \mathcal{V}^{n+k}$. **This gives**

$$h_\mu(\mathcal{V}^k, T) = \overline{\lim}_n \frac{1}{n} H_\mu(\mathcal{V}^{n+k}) = h_\mu(\mathcal{V}, T).$$

The last equation holds because $\lim_n \frac{n}{n+k} = 1$.

COROLLARY: This implies $h_\mu(\mathcal{V}, T) = \frac{1}{n} h_\mu(\mathcal{V}^n, T^n)$.

Proof: Indeed, $\bigvee_{j=0}^{kn-1} \mathcal{V}^n = \mathcal{V}^{kn^2}$, giving $h_\mu(\mathcal{V}^n, T^n) = \overline{\lim}_n \frac{1}{n} H_\mu(\mathcal{V}^{kn}) = n h_\mu(\mathcal{V}, T)$ (the last equation is implied by the previous remark). ∎

COROLLARY: For any (M, μ, T), one has $h_\mu(T^n) = n h_\mu(T)$.

Proof: Since $\mathcal{V}^n$ is a refinement of $\mathcal{V}$, one has $H_\mu(\mathcal{V}^n) \geqslant H_\mu(\mathcal{V})$. This gives $h_\mu(T^n) = \sup_{\mathcal{V}} H_\mu(T^n, \mathcal{V}) = \sup_{\mathcal{V}^n} H_\mu(T^n, \mathcal{V}^n) = n \sup_{\mathcal{V}} H_\mu(T, \mathcal{V}) = n h_\mu(T)$. ∎

COROLLARY: Let $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ be a sum of atomic measures. Since T preserves μ, T acts on the set $\{x_1, ..., x_n\}$ by permutations. **Therefore $T^{n!} = \mathrm{Id}$, giving**

$$h_\mu(\mathcal{V}, T) = h_\mu(\mathcal{V}^{n!}, T) = \frac{1}{n!} h_\mu(\mathcal{V}^{n!}, T^{n!}) = 0.$$

∎

Independent partitions

DEFINITION: Let $\mathcal{V}$, $\mathcal{W}$ be finite partitions. We say that they are **independent** if for all $V_i \in \mathcal{V}$ and $W_j \in \mathcal{W}$, one has $\mu(V_i \cap W_j) = \mu(V_i)\mu(W_j)$.

REMARK: In probabilistic terms, this means that the **events associated with V_i and W_j are uncorrelated**.

REMARK: Let $\mathcal{V}$, $\mathcal{W}$ be independent partitions, with $p_1, ..., p_k$ measures of V_i and $q_1, ..., q_l$ measures of $\mathcal{W}$. **Then**

$$H_\mu(\mathcal{V} \vee \mathcal{W}) = \sum_{i,j} p_i q_j \log(p_i q_j) = \sum_j \sum_i p_i q_j \log q_j + \sum_i \sum_j q_j p_i \log p_i = H_\mu(\mathcal{V}) + H_\mu(\mathcal{W}).$$

COROLLARY: Let (M, μ, T) be a dynamical system, and $\mathcal{V}$ a partition of M. Assume that $T^{-i}(\mathcal{V})$ is independent from $\mathcal{V}^i$ for all i. **Then $H_\mu(\mathcal{V}^n) = nH_\mu(\mathcal{V})$, giving $h_\mu(T, \mathcal{V}) = H_\mu(\mathcal{V})$.**

REMARK: It is possible to show (and it clearly follows from Shannon's description of entropy) that $H(\mathcal{V} \vee \mathcal{W}) \leqslant H(\mathcal{V}) + H(\mathcal{W})$, **and the equality is reached if and only if $\mathcal{V}$ and $\mathcal{W}$ are independent.** This result is called **subadditivity of entropy**. This implies, in particular, that $H_\mu(\mathcal{V}^n) \leqslant nH_\mu(\mathcal{V})$, hence **the limit $\lim \frac{1}{n} H_\mu(\mathcal{V}^n)$ is always finite.**

Entropy of dynamical system: Bernoulli space

DEFINITION: Let P be a finite set, $P^{\mathbb{Z}}$ the product of $\mathbb{Z}$ copies of P, $\Sigma \subset \mathbb{Z}$ a finite subset, and $\pi_\Sigma: \ P^{\mathbb{Z}} \longrightarrow P^{|\Sigma|}$ projection to the corresponding components. **Cylindrical sets** are sets $C_R := \pi_\Sigma^{-1}(R)$, where $R \subset P^{|\Sigma|}$ is any subset.

REMARK: For Bernoulli space, **a complement to an cylindrical set is again a cylindrical set,** and the cylindrical sets **form a Boolean algebra.**

DEFINITION: Bernoulli measure on $P^{\mathbb{Z}}$ is μ such that $\mu(C_R) := \frac{|R|}{|P|^{|\Sigma|}}$.

EXAMPLE: Let $\mathcal{V} = \{V_i\}$ be a finite partition of Bernoulli space $M = P^{\mathbb{Z}}$ into cylindrical sets, a T the Bernoulli shift. Let $\Sigma \subset \mathbb{Z}$ be a finite subset such that all V_i are obtained as $\pi_\Sigma^{-1}(R_i)$ for some $R_i \subset P^{|\Sigma|}$. For N sufficienty big, the sets Σ and $T^{-i}(\Sigma)$ don't intersect. In this case, **the partitions $\mathcal{V}^{kN}$ and $T^{-N}(\mathcal{V})$ are independent, giving $h_\mu(T^N, \mathcal{V}) = H_\mu(\mathcal{V})$.** Since $h_\mu(T) = 1/N h_\mu(T^N) \geqslant H_\mu(\mathcal{V})$, **this implies that the entropy of T is positive.**

Approximating partitions

LEMMA 1: Let (M,μ) be a space with measure, and A an algebra of measurable subsets of M which generates any measurable subset uo to measure 0. Then for any partition $\mathcal{V}$ with finite entropy and any $\varepsilon 0$. **there exists a finite partition $\mathcal{W} \subset A$ such that $H_\mu(\mathcal{W} \vee \mathcal{V}) - H_\mu(\mathcal{W}) < \varepsilon$.**

Proof: Using Lebesgue approximation theorem, we can approximate the partition $\mathcal{V}$ by $\mathcal{W} \subset A$ with arbitrary precision: for each $V_i \in \mathcal{V}$ there exists $W_i \in \mathcal{W}$ (which can be empty) such that $\mu(V_i \triangle W_i) < \varepsilon_i$. Then

$$H_\mu(\mathcal{W} \vee \mathcal{V}) - H_\mu(\mathcal{W}) = \sum_i p_i H_\mu(p_i^{-1}\mu(W_i \cap V_1), ..., p_i^{-1}\mu(W_i \cap V_n)).$$

where $p_i = \mu(W_i)$. However, $\mathcal{W}$ is chosen in such a way that $\mu(W_i \cap V_i)$ is arbitrarily close to p_i, and $\mu(W_i \cap V_j)$ is arbitrarily small for $j \neq i$, hence the entropy $H_\mu(p_i^{-1}\mu(W_i \cap V_1), ..., p_i^{-1}\mu(W_i \cap V_n))$ is arbitrarily small. ■

Kolmogorov-Sinai theorem

THEOREM: (Kolmogorov-Sinai)
Let (M, μ, T) be a dynamical system, and $\mathcal{V}_1 \prec \mathcal{V}_2 \prec ...$ a sequence of partitions of M finite entropy, such that the subsets $\bigcup_{i=1}^{\infty} \mathcal{V}_i$ generate the σ-algebra of measurable sets, up to measure zero. **Then** $h_\mu(T) = \lim_n h_\mu(T, \mathcal{V}_n)$.

Proof: Notice that $h_\mu(T, \mathcal{V}_n)$ is monotonous as a function of n, because $\mathcal{V}_1 \prec \mathcal{V}_2 \prec ....$ Moreover, $h_\mu(T, \mathcal{V}_n^N) = h_\mu(T, \mathcal{V}_n)$ as shown above. Since any partition $\mathcal{W}$ admits an approximation by a partition from the σ-algebra generated by $\mathcal{V}_n$, we obtain that for n sufficiently big, one has $h_\mu(T, \mathcal{W}) \leqslant h_\mu(T, \mathcal{V}_n^N) + \varepsilon = h_\mu(T, \mathcal{V}_n) + \varepsilon$ Passing to the limit as $\varepsilon \longrightarrow 0$, obtain that $h_\mu(T, \mathcal{W}) \leqslant \lim_n h_\mu(T, \mathcal{V}_n)$. ■

DEFINITION: We say that a partition $\mathcal{V}$ is **a generator**, or **generating partition** if the union of all $\mathcal{V}^n = \bigvee_{i=0}^{n-1} T^{-i}(\mathcal{V})$ generates the σ-algebra of measurable sets, up to measure zero.

COROLLARY: Let $\mathcal{V}$ be a generating partition on (M, μ, T). **Then** $h_\mu(T) = h_\mu(T, \mathcal{V})$.

Proof: By Kolmogorov-Sinai, $h_\mu(T) = \lim_n h_\mu(T, \mathcal{V}^n)$. However, $h_\mu(T, \mathcal{V}^n) = h_\mu(T, \mathcal{V})$ as shown above. ■

Entropy of a dynamical system: Bernoulli space (2)

REMARK: Let $(M = P^{\mathbb{Z}}, \mu, T)$ be the Bernoulli system, with $P = \{x_1, ..., x_p\}$ and Π_i the projection to i-th component. Consider a partition $\mathcal{V}$ with $M = \coprod_{i=1}^{p} \Pi_0^{-1}(x_i)$. Clearly, the Borel σ-algebra is generated by $\Pi_i^{-1}(\{x\})$. Then $\mathcal{V}$ is a generating partition. However, $h_\mu(T, \mathcal{V}) = \sum_{i=1}^{p} \frac{1}{[} \log(p) = \log(p)$. **We have proved that $h_\mu(T) = \log(|P|)$.**

Entropy and measure decomposition

PROPOSITION: Let M be a space with σ-algebra, T a measurable map, $t \in [0,1]$ and μ, ν be T-invariant measures. Consider the measure $\rho := t\mu + (1-t)\nu$. **Then** $h_\rho(T, \mathcal{V}) = th_\mu(T, \mathcal{V}) + (1-t)h_\rho(T, \mathcal{V})$.

Proof. Step 1: For any $p_1, ..., p_n, q_1, ..., q_n \in [0,1]$ with $\sum q_i = \sum p_i = 1$, we have

$$-\sum_i (tp_i + (1-t)q_i) \log(tp_i + (1-t)q_i) \geqslant -t \sum_i p_i \log p_i - (1-t) \sum_i q_i \log q_i, \quad (*)$$

because the function $x \mapsto -x \log x$ is concave. On the other hand, $-\log(tp_i + (1-t)q_i) \leqslant -\log(tp_i)$, because $x \mapsto -\log x$ is monotonously decreasing. This gives

$$-\sum_i (tp_i + (1-t)q_i) \log(tp_i + (1-t)q_i) \leqslant -\sum_i tp_i \log(tp_i) - \sum_i tq_i \log((1-t)q_i) =$$
$$-t \sum_i p_i \log p_i - (1-t) \sum_i q_i \log q_i - \sum_i p_i t \log t - \sum_i p_i (1-t) \log(1-t). \quad (**)$$

The last two terms of (**) give

$$-\sum_i p_i t \log t - \sum_i p_i (1-t) \log(1-t) = -t \log t - (1-t) \log(1-t),$$

because $\sum q_i = \sum p_i = 1$.

Entropy and measure decomposition (2)

Proof. Step 1: For any $p_1, ..., p_n, q_1, ..., q_n \in [0,1]$ with $\sum q_i = \sum p_i = 1$, we have

$$-\sum_i (tp_i + (1-t)q_i)\log(tp_i + (1-t)q_i) \geqslant -t\sum_i p_i \log p_i - (1-t)\sum_i q_i \log q_i, \quad (*)$$

$$-\sum_i (tp_i + (1-t)q_i)\log(tp_i + (1-t)q_i) \leqslant -t\sum_i p_i \log p_i - (1-t)\sum_i q_i \log q_i - \\ -t\log t - (1-t)\log(1-t) \quad (**)$$

Step 2: Comparing the inequalities (*) and (**), we obtain

$$tH_\mu(\mathcal{V}) + (1-t)H_\nu(\mathcal{V}) \leqslant H_\rho(\mathcal{V}) \leqslant tH_\mu(\mathcal{V}) + (1-t)H_\nu(\mathcal{V}) - t\log t - (1-t)\log(1-t)$$

Passing to the limit of $\frac{1}{n}H(\mathcal{V}^n)$ and using $\lim_n \frac{1}{n}(-t\log t - (1-t)\log(1-t)) = 0$. we obtain that $h_\rho(T,\mathcal{V}) = th_\mu(T,\mathcal{V}) + (1-t)h_\nu(T,\mathcal{V})$. ■

Jacobs theorem

REMARK: We have just shown that **entropy of a partition is affine under finite linear combination of probability measures.** However, **this statement is false** for a continuous decomposition of measures. Indeed, the **entropy of a partition is not continuous in the weak topology on measures**. For example, **entropy vanishes on all measures with finite support, but any Radon measure is a limit of measures with finite support.**

However, **the entropy of a dynamical system is affine under the ergodic decomposition.**

The proof of the following theorem will be omitted.

THEOREM: (K. Jacobs)
Let (M, μ, T) be a dynamical system, with M a complete metric space with countable base. Let E be the set of all ergodic measures, and consider the ergodic decomposition $\mu = \int_E \nu\kappa$, where $\nu \in E$ and κ is the corresponding measure on E (its existence and uniqueness we proved in Lecture 19). **Then $h_\mu(T) = \int_E h_\nu(T)\kappa$.**

Topological entropy

DEFINITION: Let M be a compact topological space, and $\{U_i \subset M\}$ an open cover, $\bigcup U_i = M$. A cover $\{V_i \subset M\}$ is called **a subcover** if it is a subset which is still a cover. Given a cover α, denote by $N(\alpha)$ the smallest cardinality of a subcover of α. **The entropy** of a cover is $H(\alpha) = \log N(\alpha)$.

DEFINITION: Let $f: M \longrightarrow M$ be a continuous map, α a cover, and $\alpha^n := \alpha \vee f^{-1}(\alpha) \vee ... \vee f^{-n+1}(\alpha)$. Define **entropy** of a map with respect to the cover by $H(f, \alpha) := \lim_n \frac{1}{n} H(\alpha^n)$.

EXERCISE: Prove that the function $n \longrightarrow H(\alpha^n)$ is **subadditive**, that is, $H(\alpha^{m+n}) \leqslant H(\alpha^m) + H(\alpha^n)$.

REMARK: For a subadditive monotonously non-decreasing sequence $\{a_i\}$, **the sequence $\frac{1}{n} a_n$ is monotonously non-increasing, hence the limit $\lim_n \frac{1}{n} a_n$ exists.** Indeed, for such sequence, $a_n - a_{n-1} > a_{n+1} - a_n$, hence $b_i := a_{n+1} - a_n$ is non-negative and monotonous, and its Cesáro sum $\frac{1}{n} a_n = \frac{1}{n} \sum_{i=1}^n b_i$ is convergent.

REMARK: The measure entropy is also subadditive, which explains convergence.

DEFINITION: Define **the topological entropy** $h(f)$ as $\sup_\alpha H(f, \alpha)$.

Metric entropy

REMARK: In old literature, "metric entropy" refers to the measure entropy defined above, and both notions of "topological entropy" (previous slide) and metric entropy (this slide) are called "topological entropy".

DEFINITION: Let $X \subset M$ be a subset of a metric space. We denote by $X(\varepsilon)$ the set $\{y \in M \mid d(y, X) < \varepsilon\}$. This set is called **$\varepsilon$-neighbourhood** of X. **An ε-net** is a subset $X \subset M$ such that $X(\varepsilon) = M$. Denote by $N(M, \varepsilon)$ the cardinality of the smallest ε-net.

DEFINITION: Let $T: M \longrightarrow M$ be a continuous map of compact metric spaces. Consider M^n as a metric space with the metric $d((x_1, ..., x_n), (y_1, ..., y_n)) = \max(d(x_1, y_1), d(x_2, y_2), ...d(x_n, y_n))$, and let $S_n := \{(x, T(x), T^2(x), ..., T^{n-1}(x)) \subset M^n\}$. Consider the number $h(T, \varepsilon) = \overline{\lim}_n \frac{1}{n} \log N(S_n, \varepsilon)$. We define **metric entropy** of T as $h(T) := \lim_{\varepsilon \to 0} h(T, \varepsilon)$.

Metric entropy, topological entropy and measure entropy

We omit the proof of the following two theorems.

THEOREM: **Metric entropy is equal to the topological entropy.**

THEOREM: For any continuous map $T: M \longrightarrow M$ of compact metric spaces, consider the number $\sup_\mu h_\mu(T)$, where $h_\mu(T)$ is measure entropy, and supremum is taken over all T-invariant probabilistic Borel measures. Then $\sup_\mu h_\mu(T) = h(T)$: **topological entropy is the supremum of measure entropy.**