Dynamics and time series: theory and applications

Stefano Marmi – Giulio Tiozzo

Scuola Normale Superiore

Lecture 5, Jan 27, 2010

Topological entropy

Topological entropy represents the exponential growth rate of the number of orbit segments which are distinguishable with an arbitrarily high but finite precision. It is invariant under topological conjugacy. Here the phase space is supposed to be a compact metric space (X,d)

Definition 4.1 *Let $S \subset X$, $n \in N$ and $\varepsilon > 0$. S is a (n, ε)–*spanning set *if for every $x \in X$ there exists $y \in S$ such that $d(f^j(x), f^j(y)) \leq \varepsilon$ for all $0 \leq j \leq n$.*

$$h_{top}(f) = \lim_{\varepsilon \to 0} \limsup_{n \to +\infty} \frac{1}{n} \log r(n, \varepsilon)$$

Here r(n, ε) is the minimal cardinality of a (n, ε)-spanning set

Metric entropy of a partition

Thus the entropy associated to the experiment is

$$H = -\sum_{i=1}^{k} p_i log\, p_i$$

$$\mathcal{P} = \{E_1, \ldots, E_k\}$$

$$p_i = \mu(E_i)$$

In view of the definition of information = - log (probability), entropy is simply the expectation of information

Metric entropy of a dynamical system (Kolmogorov-Sinai entropy)

Given two partitions $\mathcal{P}$ and $\mathcal{Q}$

$\mathcal{P} \vee \mathcal{Q}$ the **join** of $\mathcal{P}$ and $\mathcal{Q}$

$B \cap C$ where $B \in \mathcal{Q}$ and $C \in \mathcal{Q}$

$T : X \to X$ measure preserving

$$\mathcal{P}_n = \mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-(n-1)}\mathcal{P}$$

$$h(T, \mathcal{P}) = \lim_{n \to \infty} \frac{1}{n} H(\mathcal{P}_n) \qquad h(T) = \sup_{\mathcal{P}} h(T, \mathcal{P})$$

Entropy of Bernoulli schemes

Let $N \geq 2$, $\Sigma_N = \{1, \dots N\}^{\mathbb{Z}}$.

$d(x,y) = 2^{-a(x,y)}$ where $a(x,y) = \inf\{|n|\,,\, n \in \mathbb{Z}\,,\, x_n \neq y_n\}$

shift $\sigma : \Sigma_N \to \Sigma_N$ $\quad \sigma((x_i)_{i\in\mathbb{Z}}) = (x_{i+1})_{i\in\mathbb{Z}}$

The topological entropy of (Σ_N, σ) is $\log N$

$(p_1, \dots, p_N) \in \Delta^{(N)}$ $\quad \nu(\{i\}) = p_i$

Definition 4.26 *The* Bernoulli scheme $BS(p_1, \dots, p_N)$ *is the measurable dynamical system given by the shift map* $\sigma : \Sigma_N \to \Sigma_N$ *with the (product) probability measure* $\mu = \nu^{\mathbb{Z}}$ *on* Σ_N.

Proposition 4.27 *The Kolmogorov–Sinai entropy of the Bernoulli scheme* $BS(p_1, \dots, p_N)$ *is* $-\sum_{i=1}^{N} p_i \log p_i$.

The variational principle

Let T be a continuous map on X compact Hausdorff: then

$$h_{top}(T) = \sup_{\mu \in M(X,T)} h_{\mu}(T)$$

(the sup is taken over all invariant Borel probability measures μ)

Example: Bernoulli schemes

$$h_{\mu}(T) = -\sum_{i} p_i \log p_i \leq \log N = h_{top}(T)$$

Remark: the sup need not be achieved (Gurevich, 1969)

Asymptotic equipartition property

Suppose that $\mathcal{P}$ is a finite generating partition of X. For every $\varepsilon > 0$ and $n \geq 1$ there exist subsets in $\mathcal{P}_n$, which are called (n, ε)-typical subsets, satisfying the following:

(i) for every typical subset $\mathcal{P}_n(x)$,

$$2^{-n(h+\varepsilon)} < \mu(\mathcal{P}_n(x)) < 2^{-n(h-\varepsilon)},$$

(ii) the union of all (n, ε)-typical subsets has measure greater than $1 - \varepsilon$, and
(iii) the number of (n, ε)-typical subsets is between $(1-\varepsilon)2^{n(h-\varepsilon)}$ and $2^{n(h+\varepsilon)}$.

These formulas assume that the entropy is measured in bits, i.e. using the base 2 logarithm

Topological Markov chains or subshifts of finite type

$$\Sigma_A = \{x \in \Sigma_N\,,\, (x_i, x_{i+1}) \in \Gamma\, \forall i \in \mathbb{Z}\} \qquad \Gamma \subset \{1, \ldots N\}^2$$

Σ_A is a compact shift invariant subset of Σ_N

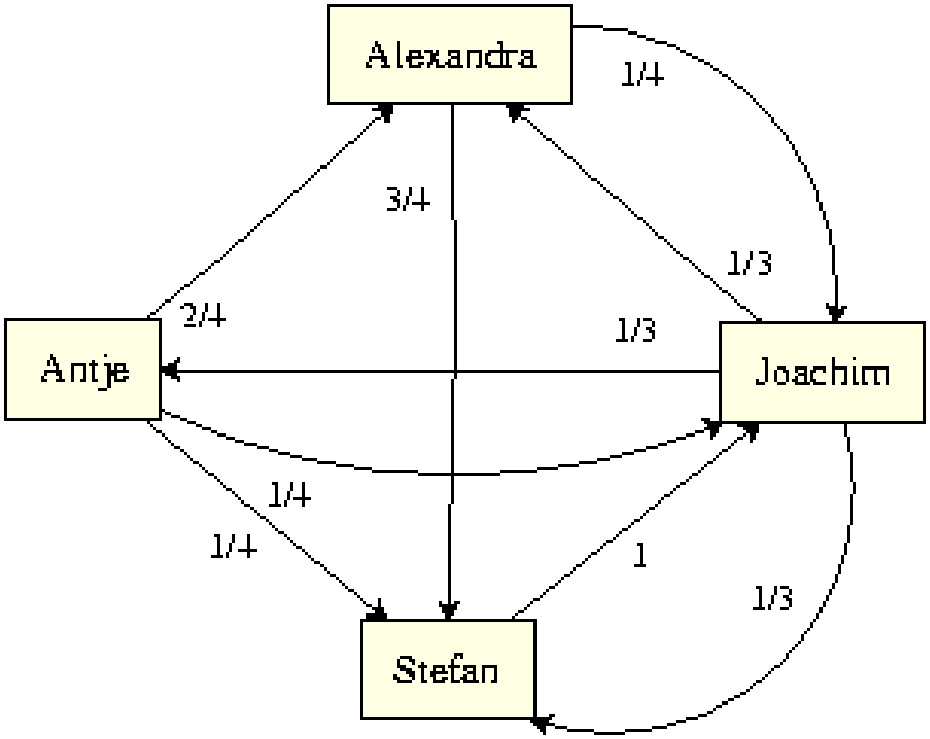
$$A = A_\Gamma \text{ the } N \times N \text{ matrix with entries } a_{ij} \in \{0, 1\}$$

$$a_{ij} = \begin{cases} 1 & \Longleftrightarrow (i,j) \in \Gamma \\ 0 & \text{otherwise} \end{cases}$$

The restriction of the shift σ to Σ_A is denoted σ_A

$A^m = (a_{ij}^m)$ and $a_{ij}^m > 0$ for all i, j (primitive matrix)

Markov chains



Topological: some moves are allowed and some are not

Metric: any allowed move happens with some fixed probability

Entropy of Markov chains

Theorem 4.35 (Perron–Frobenius, see [Gan]) If A is primitive then there exists an eigenvalue $\lambda_A > 0$ such that :

(i) $|\lambda_A| > \lambda$ for all eigenvalues $\lambda \neq \lambda_A$;

(ii) the left and right eigenvectors associated to λ_A are strictly positive and are unique up to constant multiples ;

(iii) λ_A is a simple root of the characteristic polynomial of A.

the topological entropy of σ_A is $\log \lambda_A$ (clearly $\lambda_A > 1$ since all the integers $a_{ij}^m > 0$)

Let $P = (P_{ij})$ be an $N \times N$ matrix such that

(i) $P_{ij} \geq 0$ for all i, j, and $P_{ij} > 0 \iff a_{ij} = 1$;

(ii) $\sum_{j=1}^{N} P_{ij} = 1$ for all $i = 1, \ldots, N$;

(iii) P^m has all its entries strictly positive.

Such a matrix is called a *stochastic matrix.* Applying Perron–Frobenius theorem to P we see that 1 is a simple eigenvalue of P and there exists a normalized eigenvector $p = (p_1, \ldots, p_N) \in \Delta^{(N)}$ such that $p_i > 0$ for all i and

$$\sum_{i=1}^{N} p_i P_{ij} = p_j \ , \quad \forall\, 1 \leq i \leq N \ .$$

We define a probability measure μ on Σ_A corresponding to P prescribing its value on the cylinders :

$$\mu\left(C\begin{pmatrix} j_0, \ldots, j_k \\ i, \ldots, i+k \end{pmatrix}\right) = p_{j_0} P_{j_0 j_1} \cdots P_{j_{k-1} j_k} \ ,$$

for all $i \in \mathbb{Z}$, $k \geq 0$ and $j_0, \ldots, j_k \in \{1, \ldots, N\}$. It is called the *Markov measure* associated to the stochastic matrix P.

: the subshift σ_A preserves the Markov measure μ.

$$h_\mu(\sigma_A) = -\sum_{i,j=1}^{N} p_i P_{ij} \log P_{ij} \qquad\qquad h_\mu(\sigma_A) \leq h_{top}(\sigma_A)$$

The entropy of English

Is English a stationary ergodic process? Probably not!

Stochastic approximations to English: as we increase the complexity of the model, we can generate text that looks like English. The stochastic models can be used to compress English text. The better the stochastic approximation, the better the compression.

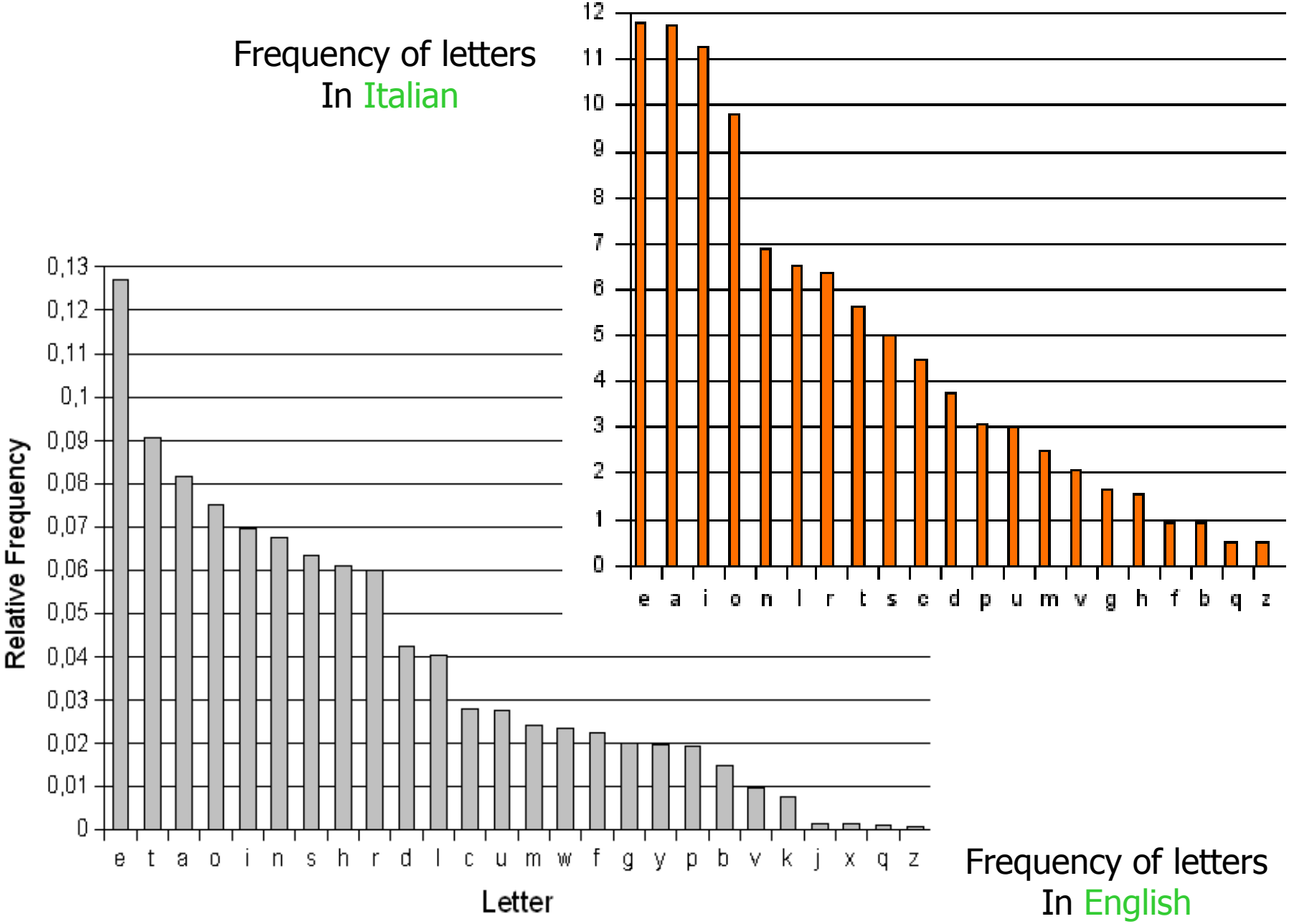
alphabet of English = 26 letters and the space symbol

models for English are constructed using empirical distributions collected from samples of text.

E is most common, with a frequency of about 13%,

least common letters, Q and Z, have a frequency of about 0.1%.

Frequency of letters
In Italian



Frequency of letters
In English

Source: Wikipedia

Construction of a Markov model for English

The frequency of pairs of letters is also far from uniform:
Q is always followed by a U, the most frequent pair is TH,
(frequency of about 3.7%), etc.

Proceeding this way, we can also estimate higher-order conditional probabilities and build more complex models for the language.

However, we soon run out of data. For example, to build a third-order Markov approximation, we must compute

$$P(x_n \mid x_{n-1}, x_{n-2}, x_{n-3})$$

in correspondence of $27 \times 27^3 = 531\,441$ entries for this table: we need to process millions of letters to make accurate estimates of these probabilities.

Examples

(Cover and Thomas, Elements of Information Theory, 2nd edition , Wiley 2006)

- Zero order approximation (equiprobable h=4.76 bits):

  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

- First order approximation (frequencies match):

  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI

  ALHENHTTPA OOBTTVA NAH BRL

- Second order (frequencies of pairs match): ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

- Third order (frequencies of triplets match): IN NO IST LAT WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

- Fourth order approximation (frequencies of quadruplets match, each letter depends on previous three letters; h=2.8 bits):

  THE GENERATED JOB PROVIDUAL BETTER TRANDTHE DISPLAYED CODE, ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT DO HOCK BOTHE MERG. (INSTATES CONS ERATION. NEVER ANY OF PUBLE AND TO THEORY. EVENTIAL CALLEGAND TO ELAST BENERATED IN WITH PIES AS IS WITH THE)

- First order WORD approximation (random words, frequencies match): REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.
- Second order (WORD transition probabilities match): THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Entropy, coding and data compression

- Computer file= infinitely long binary sequence
- Entropy = best possible compression ratio
- Lempel-Ziv (Compression of individual sequences via variable rate coding, IEEE Trans. Inf. Th. 24 (1978) 530-536): it does not assume knowledge of probability distribution of the source and achieves asymptotic compression ratio=entropy of source

Lempel – Ziv algorithm

Encoding Algorithm

1. Initialize the dictionary to contain all blocks of length one (E.g.: D={a,b}).
2. Search for the longest block **W** which has appeared in the dictionary.
3. Encode **W** by its index in the dictionary.
4. Add **W** followed by the first symbol of the next block to the dictionary.
5. Go to Step 2.

Data: a b b a a b b a a b a b b a a a a b a a b b a …

Source: www.data-compression.com

Encoding Algorithm

1. Initialize the dictionary to contain all blocks of length one (D={a,b}).
2. Search for the longest block **W** which has appeared in the dictionary.
3. Encode **W** by its index in the dictionary.
4. Add **W** followed by the first symbol of the next block to the dictionary.
5. Go to Step 2.

Data: a b b a a b b a a b a b b a a a a b a a b b a

Dictionary			
Index	Entry	Index	Entry
0	a	7	
1	b	8	
2		9	
3		10	
4		11	
5		12	
6		13	

Source: www.data-compression.com

Law of large numbers

$\{\mathbf{X_i}\}$ independent identically distributed random variables

$$\mathrm{E}(\mathrm{X_i}) = \mu < +\infty$$

Then $$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

Weak form:

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| < \varepsilon) = 1$$

Strong form:

$$\overline{X}_n \to \mu \qquad \text{almost surely}$$

Law of large numbers vs Birkhoff theorem

Random setting

$\{\mathbf{X_i}\}$ i.i.d. random variables

$$\mathrm{E}(\mathrm{X_i}) = \mu < +\infty$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

almost surely

Deterministic setting

$$T: X \to X$$

$f \in L^1(X, d\mu)$ observable

$$X_i := \{ f \circ T^i \}$$

are not necessarily independent

If T ergodic

$$\frac{1}{n}\sum_{i=1}^{n} f \circ T^i \to \int f \, d\mu$$

almost surely

Central limit theorem

$\{\mathbf{X_i}\}$ independent identically distributed random variables

$$\mathrm{E}(\mathrm{X_i}) = \mu < +\infty \qquad \mathrm{Var}(\mathrm{X_i}) = \sigma^2 > 0$$

$$Z_n := \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{weak} N(0,1)$$

$$\lim_{n\to\infty} \mathrm{P}(Z_n \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

Central limit theorem for deterministic systems

(X, A, μ, T) ergodic measurable dynamical system

$f \in L^2(X, d\mu)$ $\int f \, d\mu = 0$

$$\sigma^2 = \int f^2 d\mu + 2\sum_{n=1}^{\infty} \int f(f \circ T^n) \, d\mu$$

Analogously to the independent case, we would like to have

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} f \circ T^i \to N(0,1)$$

Central limit theorem for deterministic systems

(X, A, μ, T) ergodic measurable dynamical system

$f \in L^2(X, d\mu)$ $\int f \, d\mu = 0$

$$\sigma^2 = \int f^2 d\mu + 2\sum_{n=1}^{\infty} \int f(f \circ T^n) \, d\mu$$

Analogously to the independent case, we would like to have

$$P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} f \circ T^i \geq z\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} \, dt$$

Hypotheses for deterministic CLT

1. Mixing-type condition

$$c(n, f) := \sup\{\int f\,(g \circ T^n)\,d\mu : \int g^2\,d\mu = 1\}$$

$$\sum_{n=0}^{\infty} c(n, f) < +\infty$$

2. Cohomological equation has no solution

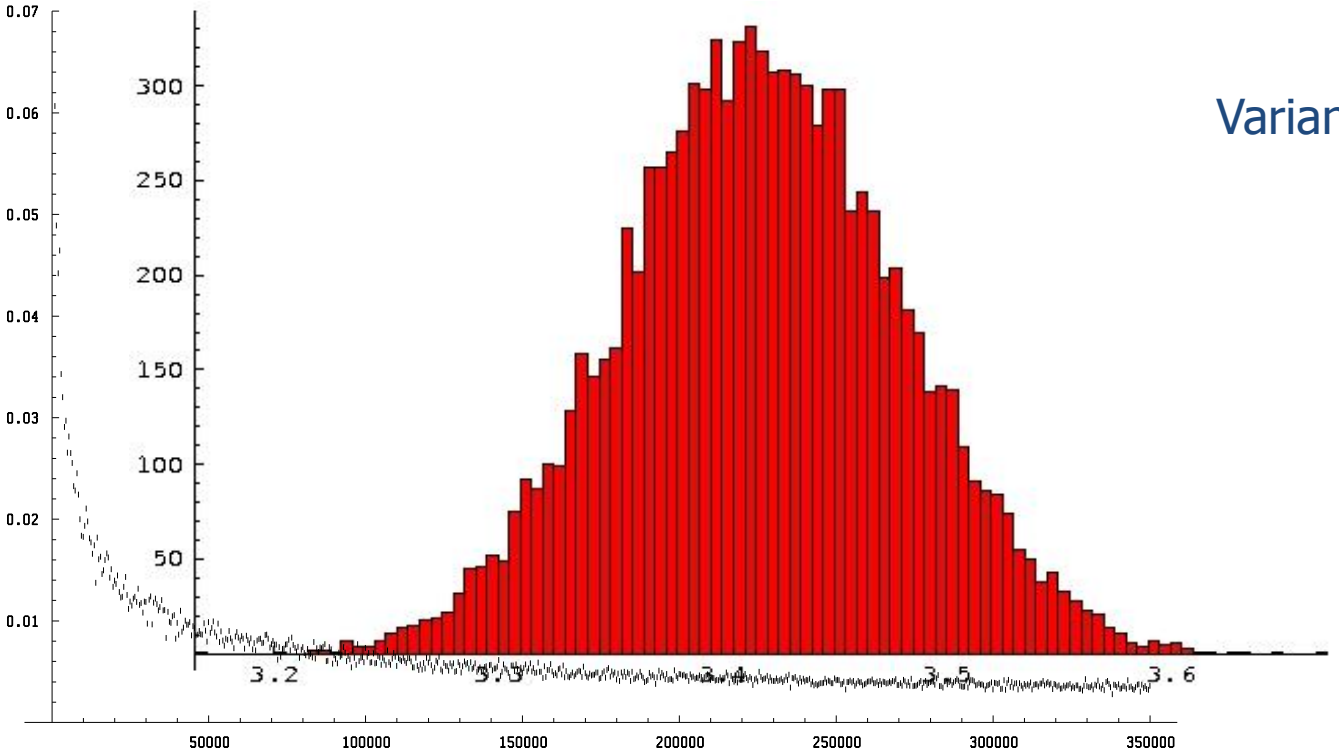
$$\sigma = 0 \Leftrightarrow \exists u \in L^2(X, \mu)\ s.t.\ f = u \circ T - u$$

Then $$P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} f \circ T^i \geq z\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\,dt$$

An example: distribution of Birkhoff averages for the entropy

$$h_\mu(T) = \int \log |T'| \, d\mu = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log |T' \circ T^i|$$

KS entropy = Birkhoff average for observable $f(x) = \log |T'(x)|$



Variance decays as

$$\frac{1}{\sqrt{n}}$$

Speed of convergence (Berry-Esseen theorem)

$\{\mathbf{X_i}\}$ independent identically distributed random variables

$$\mathrm{E}(\mathrm{X_i}) = \mu < +\infty \qquad \mathrm{Var}(\mathrm{X_i}) = \sigma^2 > 0$$

$$\mathrm{E}(|\,\mathrm{X_i}\,|^3) = \rho < +\infty$$

By CLT $$\frac{X_1 + \ldots + X_n}{\sigma\sqrt{n}} \to N(0,1)$$

Speed of convergence (Berry-Esseen theorem)

$\{\mathbf{X_i}\}$ independent identically distributed random variables

$$\mathrm{E}(X_i) = \mu < +\infty \qquad \mathrm{Var}(X_i) = \sigma^2 > 0$$

$$\mathrm{E}(|X_i|^3) = \rho < +\infty$$

By CLT $$F_n(x) := \mathrm{P}\left(\frac{X_1 + \ldots + X_n}{\sigma\sqrt{n}} \geq x \right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Moreover

$$\left| F_n(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \right| \leq \frac{0.7056\,\rho}{\sigma^3\sqrt{n}}$$