Dynamical systems, information and time series

Stefano Marmi
Scuola Normale Superiore
http://homepage.sns.it/marmi/

Lecture 2- European University Institute
September 25, 2009

- Lecture 1: An introduction to dynamical systems and to time series. Periodic and quasiperiodic motions. (Tue Jan 13, 2 pm - 4 pm Aula Bianchi)
- Lecture 2: A priori probability vs. statistics: ergodicity, uniform distribution of orbits. The analysis of return times. Kac inequality. Mixing (Sep 25)
- Lecture 3: Shannon and Kolmogorov-Sinai entropy. Randomness and deterministic chaos. Relative entropy and Kelly's betting. (Oct 9)
- Lecture 4: Time series analysis and embedology: can we distinguish deterministic chaos in a noisy environment? (Oct 30)
- Lecture 5: Fractals and multifractals. (Nov 6)

References:

- Fasano Marmi "Analytical Mechanics" Oxford University Press Chapter 13
- Benjamin Weiss: “Single Orbit Dynamics”, AMS 2000.
- Daniel Kaplan and Leon Glass: "Understanding Nonlinear Dynamics" Springer (1995)
- Sauer, Yorke, Casdagli: Embedology. J. Stat. Phys. 65 (1991) 579-616
- Michael Small "Applied Nonlinear Time Series Analysis" World Scientific
- K. Falconer: "Fractal Geometry - Mathematical Foundations and Applications" John Wiley,

The slides of all lectures will be available at my personal webpage: http://homepage.sns.it/marmi/

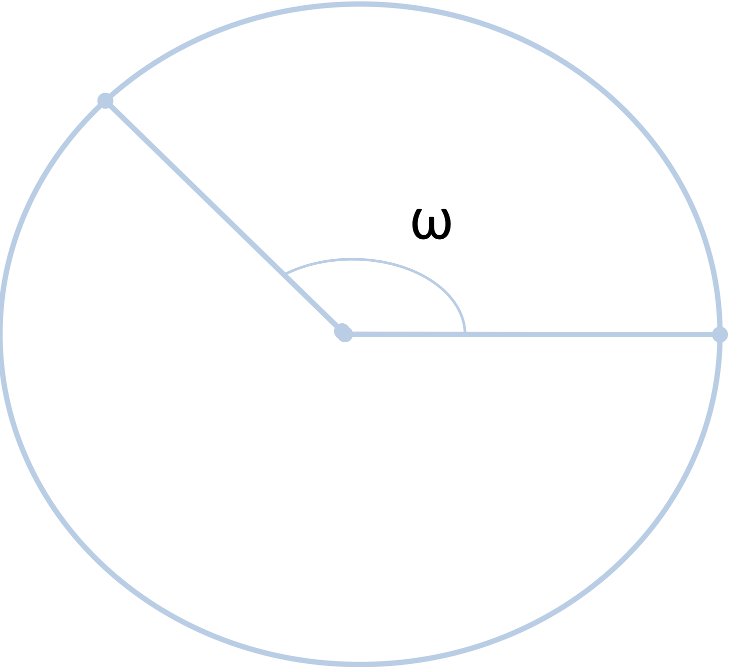
An overview of today’s lecture

- Dynamical systems
- Ergodic theorem, recurrence times
- Entropy
- Statistical induction, backtesting and black swans
- Information, uncertainty and entropy
- Risk management and information theory

Dynamical systems

- A dynamical system is a couple (phase space, time evolution law)
- The time variable can be discrete (evolution law = iteration of a map) or continuous (evolution law = flow solving a differential equation)
- The phase space is the set of all possible states (i.e. initial conditions) of our system
- Each initial condition uniquely determines the time evolution (*determinism*)
- The system evolves in time according to a fixed law (iteration of a map, differential equation, etc.)
- Often (but not necessarily) the evolution law is not linear

The simplest dynamical systems



- The phase space is the circle: **S=R/Z**

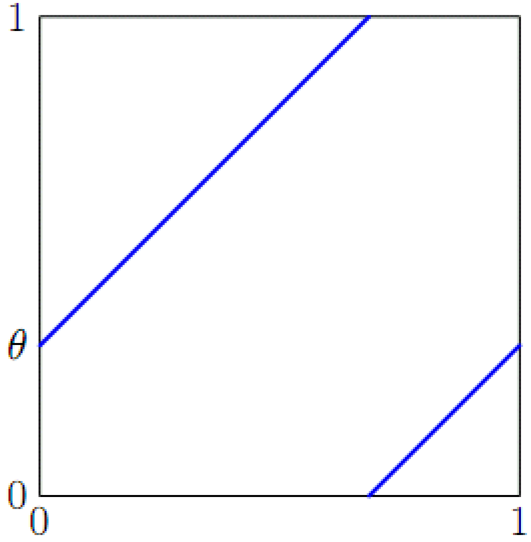
- Case 1: quasiperiodic dynamics

  $\theta(n+1)=\theta(n)+\omega \pmod 1$

  (ω irrational, for example $\theta=(\sqrt{5}-1)/2=0.618033989\ldots$)

- Case 2: chaotic dynamics

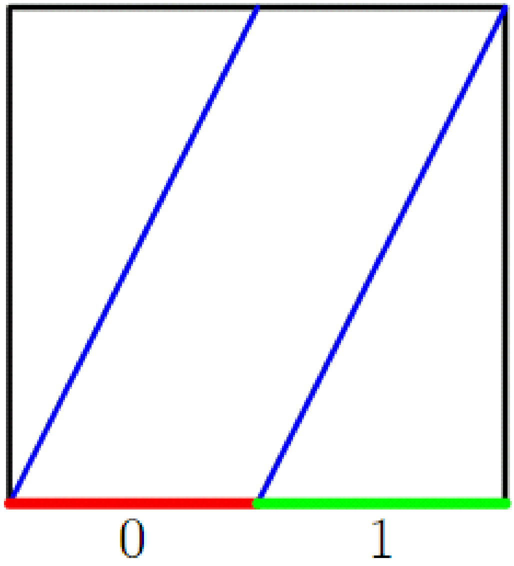
  $\theta(n+1)=2\theta(n)(\bmod 1)$

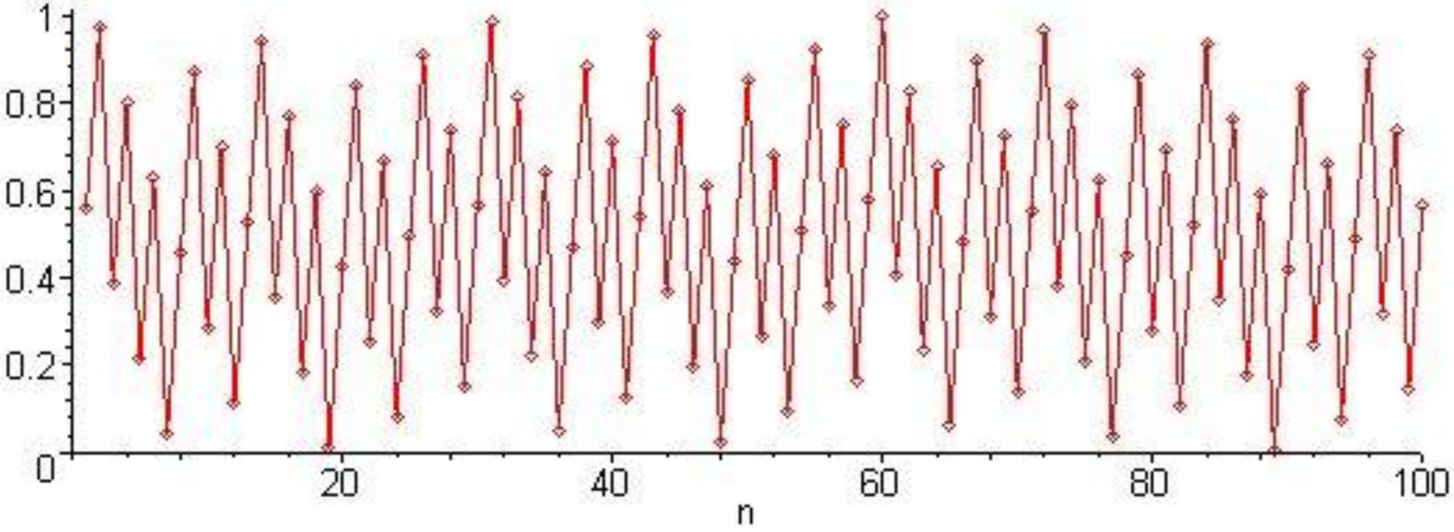


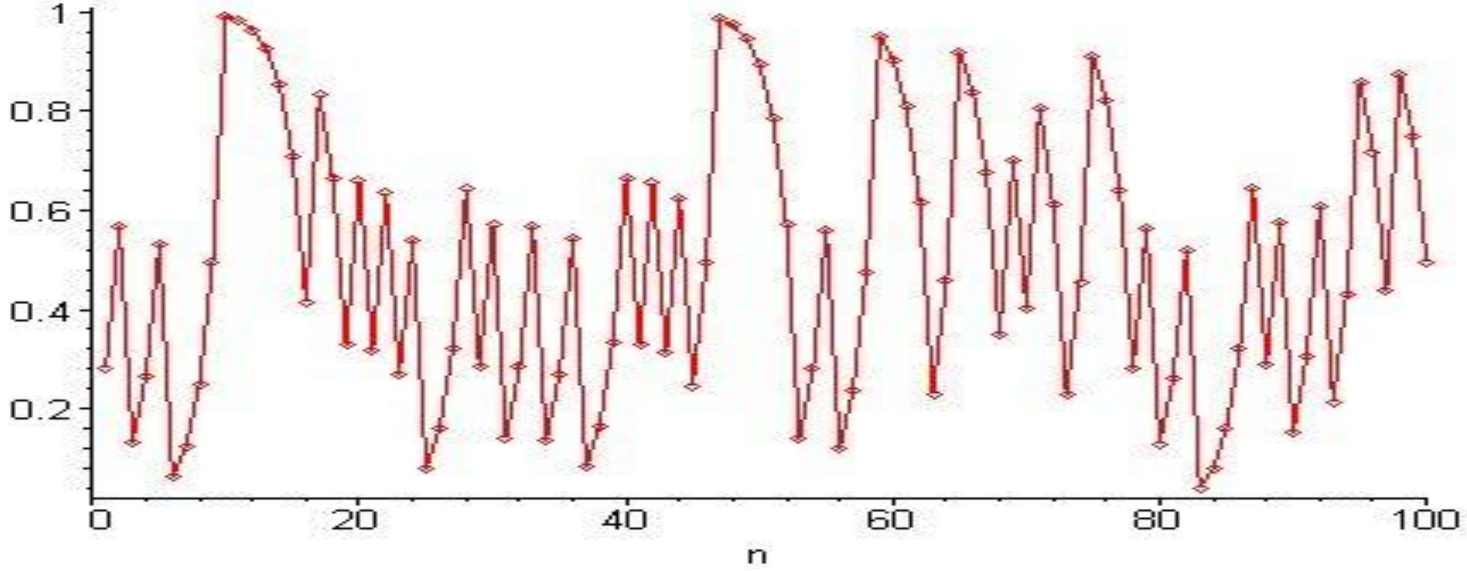
$T(\theta) = \theta + \omega \pmod 1$.

phase space $X = [0, 1)$

$T : \theta \rightarrow 2\,\theta \pmod 1$.




1
0.8
0.6
0.4
0.2
0
20
40
60
80
100
n


1
0.8
0.6
0.4
0.2
0
20
40
60
80
100
n

Sensitivity to initial conditions

For the doubling map on the circle (case 2) one has

$\theta(N)- \theta'(N)=2^N (\theta(0)- \theta'(0))$ ➡ even if the initial datum is known with a 10 digit accuracy, after 40 iterations one cannot even say if the iterates are larger than ½ or not

In quasiperiodic dynamics this does not happen: for the rotations on the circle one has $\theta(N)- \theta'(N)= \theta(0)- \theta'(0)$ and long term prediction is possible

The dynamics of the doubling maps is heterogeneous and unpredictable, quasiperiodic dynamics is homogeneous and predictable

Chaotic dynamics

Sensitive dependence on initial conditions

Density of periodic orbits

Some form of irreducibility (topological transitivity, ergodicity, etc.)

Information is produced at a positive rate: positive entropy (Lyapunov exponents)

Ergodic theory

The focus of the analysis is mainly on the *asymptotic ditribution of the orbits*, and not on transient phenomena.

Ergodic theory is an attempt to study the *statistical behaviour of orbits* of dynamical systems restricting the attention to their asymptotic distribution.

One waits until all transients have been wiped off and looks for an *invariant probability measure describing the distribution of typical orbits*.

Ergodic theory: the setup (measure preserving transformations, stationary stochastic process)

X phase space, μ probability measure on X

Φ:X → **R** **observable**, $\mu(\Phi) = \int_X \Phi \, d\mu$ **expectation value of** Φ

A measurable subset of X (**event**). A dynamics T:X→X induces a time evolution:

on observables $\Phi \rightarrow \Phi\ T$

on events $A \rightarrow T^{-1}(A)$

T **is measure-preserving** if $\mu(\Phi) = \mu(\Phi\ T)$ for all Φ, equivalently $\mu(A)=\mu(T^{-1}(A))$ for all A

Birkhoff theorem and ergodicity

Birkhoff theorem: if T preserves the measure μ then almost surely the time averages of the observables exist (statistical expectations). The system is ergodic *if these time averages do not depend on the orbit* (statistics and a-priori probability agree)

$$\frac{1}{N}\sum_{0}^{N-1}\varphi\circ T^i(x) := \frac{1}{N}S_N\varphi(x)\longrightarrow \int_X \varphi(t)\,d\mu(t)$$

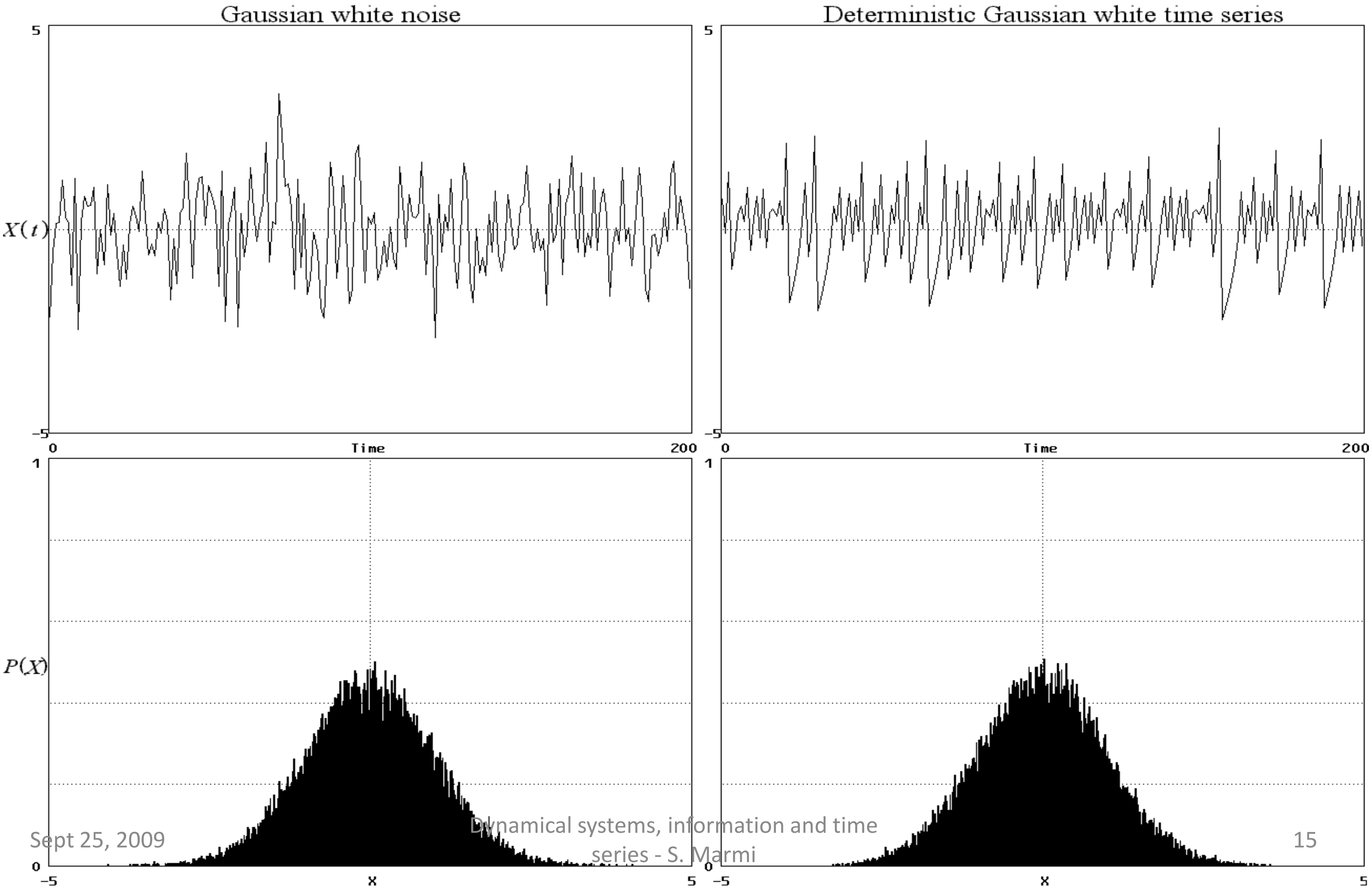
$$\frac{1}{N}\#\left\{i\in[0,N), T^i(x)\in A\right\}\longrightarrow \mu(A)$$

Law of large numbers:
Statistics of orbits =
a-priori probability

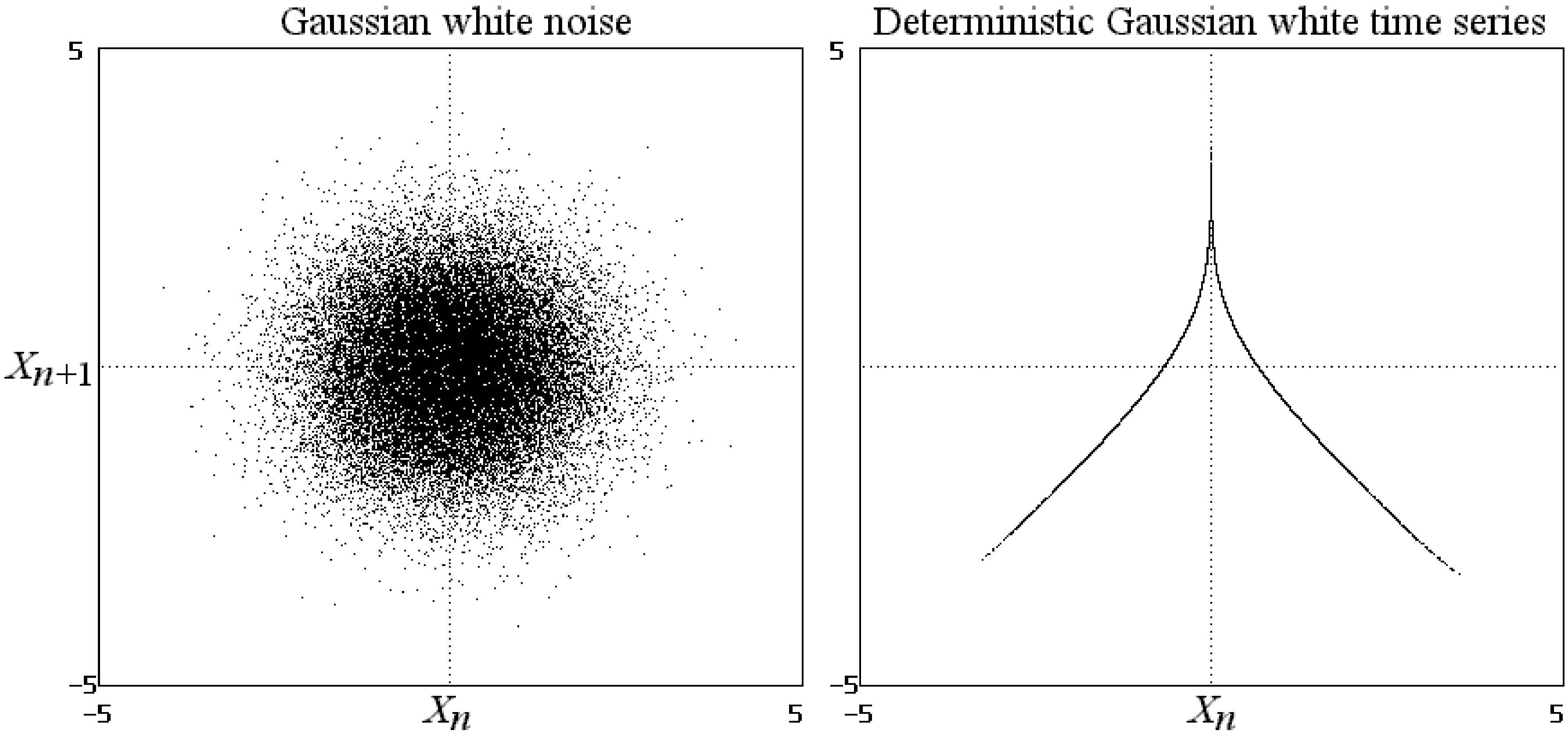
Stochastic or chaotic?

- An important goal of time-series analysis is to determine, given a times series (e.g. HRV) if the underlying dynamics (the heart) is:
  - Intrinsically *random*
  - Generated by a *deterministic nonlinear chaotic system* which generates a random output
  - *A mix of the two* (stochastic perturbations of deterministic dynamics)

Deterministic or truly random?



Time delay map



Dynamics, probability, statistics and the problem of induction

- The probability of an event (when it exists) is almost always impossible to be known a-priori
- The only possibility is to replace it with the frequencies measured by observing how often the event occurs
- The problem of *backtesting*
- The problem of *ergodicity* and of *typical points*: from a single series of observations I would like to be able to deduce the invariant probability
- *Bertrand Russell's chicken* (turkey nella versione USA)

Bertrand Russel
(The Problems of Philosophy,
Home University Library, 1912. Chapter VI On Induction) Available at the page
http://www.ditext.com/russell/rus6.html

Domestic animals expect food when they see the person who feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.

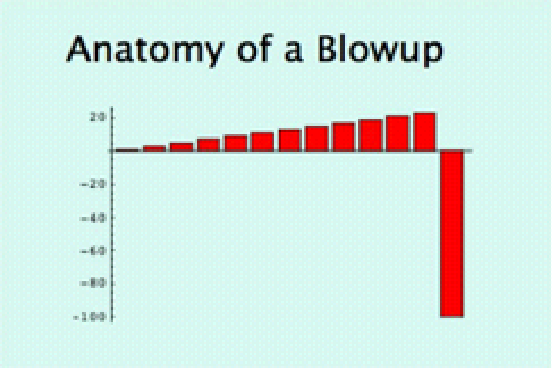


Figure 1 *My classical metaphor: A Turkey is fed for a 1000 days—every days confirms to its statistical department that the human race cares about its welfare "with increased statistical significance". On the 1001st day, the turkey has a surprise.*

http://www.edge.org/3rd_culture/taleb08/taleb08_index.html

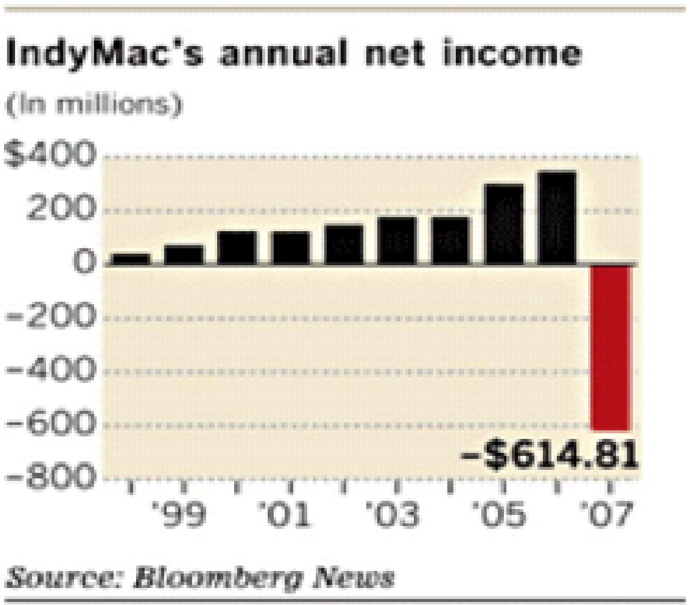


Figure 2 *The graph above shows the fate of close to 1000 financial institutions (includes busts such as FNMA, Bear Stearns, Northern Rock, Lehman Brothers, etc.). The banking system (betting AGAINST rare events) just lost > 1 Trillion dollars (so far) on a single error, more than was ever earned in the history*

http://www.edge.org/3rd_culture/taleb08/taleb08_index.html

Historical behaviour: what if the time average did not exist?

Kolakoski automatic sequence (1965): start with the digit 2. The rule is: the sequence of lengths of consecutive 1's or 2's in the sequence is the same as the sequence itself:

2

22

2211

221121

221121221

22112122122112…

An open problem

2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 2, 1, 2, 2 …

We do not know if the density of 1's exists and it is equal to =1/2 as it is conjectured on the basis of numerical simulations

The sequence can be generated starting with 22 and applying the block-substitution rules 22 →2211, 21 → 221, 12 → 211, 11 → 21 (Lagarias) thus its algorithmic complexity is very low

Entropy also vanishes, thus this is not a very random sequence n Funds - Hist… St. Louis Fed: Econo…

bri BETA entropy kolakoski sequence Cerca nei libri

libro Anteprima libro **The Mathematics of Long-Range Aperiodic Order** Di R. V. Moody, North Atlantic Treaty

Pagina 117 Cerc

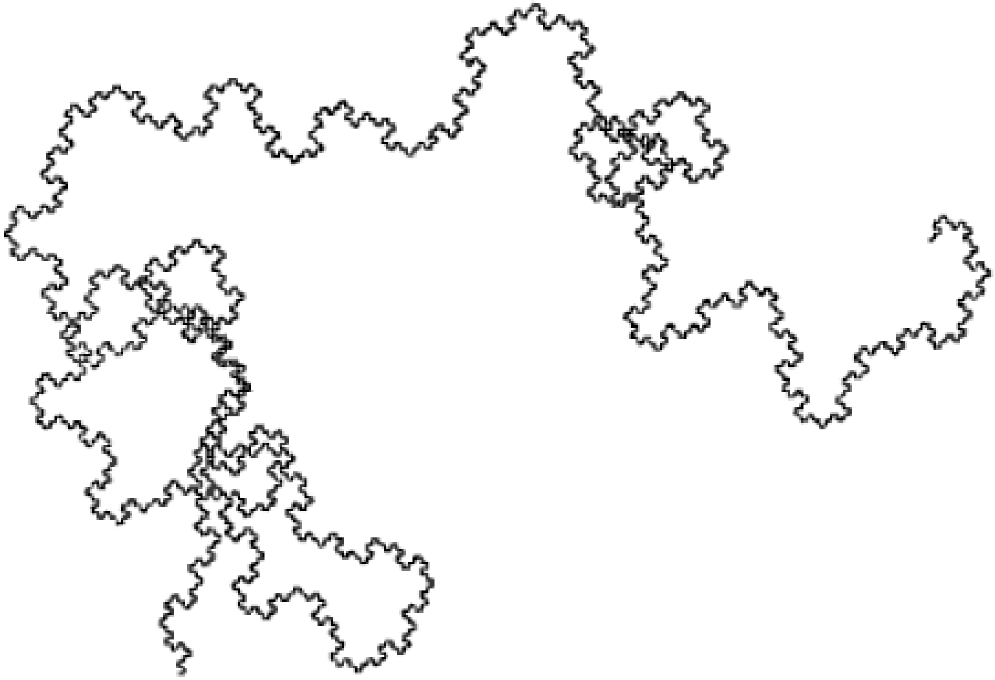


Figure 2. A walk in the plane generated by the Kolakoski sequence

of D1L-systems). The ***2-block substitution*** σ which generates x is given by

In th
mat
orde
and
Scr
Agg
Indi

Ord

Spr
Am
Hoe
iBS
Libr

Trov
Trov

Link

Min
Sm
Ont

Recurrence times

- A point is recurrent when it is a point of accumulation of its future (and past) orbit
- Poincarè recurrence: given a dynamical system T which preserves a probability measure μ and a set of positive measure E a point x of E is almost surely recurrent
- First return time of x in E:

$$R(x,E)=\min\{n>0,\ T^n x \in E\}$$

- E could be an element of a partition of the phase space (symbolic dynamics): this point of view is very important in applications (e.g. the proof of optimality of the Lempel-Ziv data compression algorithm)

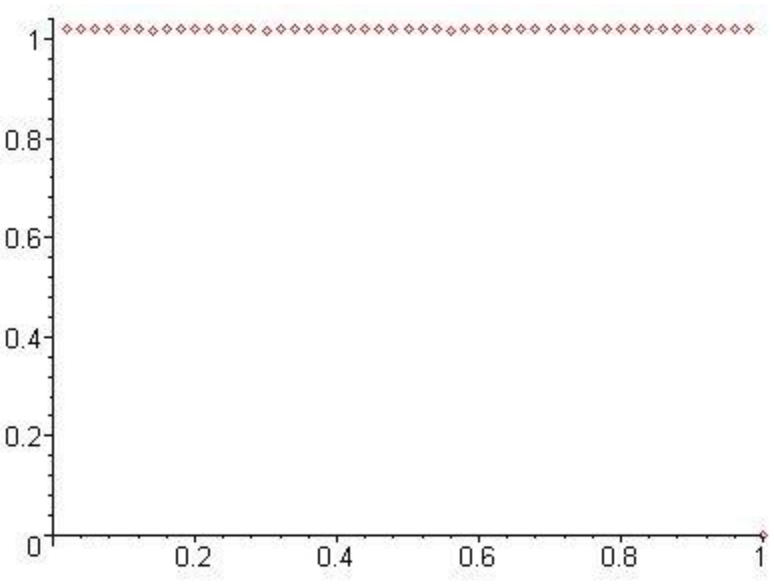
Kac’s Lemma

- If T is ergodic and E has positive measure then

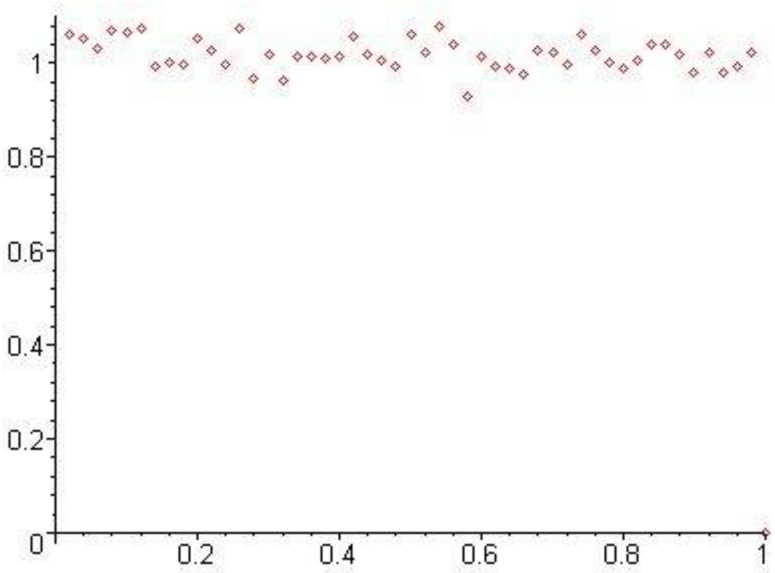
  $\int_E R(x,E)d\mu(x)=1$,

  i.e. R(x,E) is of the order of $1/\mu(E)$: the average length of time that you need to wait to see a particular symbol is the reciprocal of the probability of a symbol. Thus, we are likely to see the high-probability strings within the window and encode these strings efficiently.

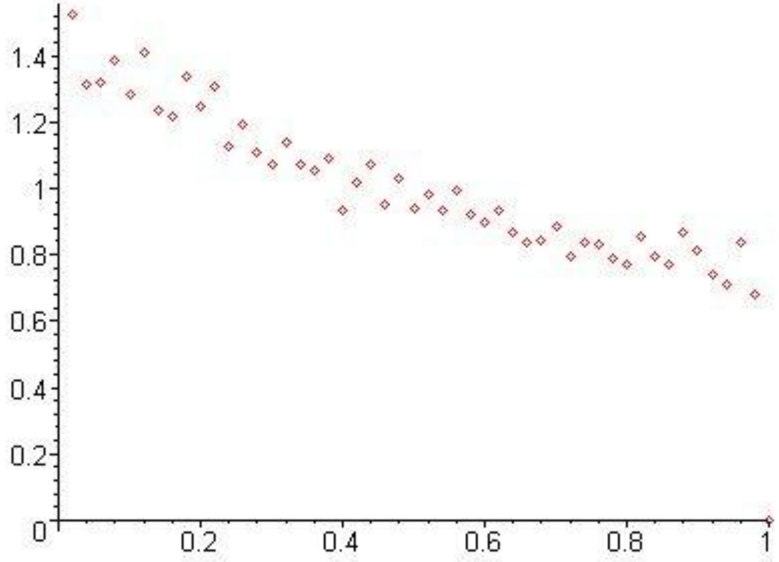
Statistical distribution of frequencies of vists



Rotation



Doubling map



Gauss map

The ubiquity of “cycles” (as long as they last…)

Furstenberg’s recurrence: If E is a set of positive measure in a measure-preserving system, and k is a positive integer, then there are infinitely many integers n for which

$$\mu(E \cap T^{-n}(E) \cap \ldots \cap T^{-(k-1)n}(E) > 0$$

- There are few persons, even among the calmest thinkers, who have not occasionally been startled into a vague yet thrilling half-credence in the supernatural, by *coincidences* of so seemingly marvellous a character that, as *mere* coincidences, the intellect has been unable to receive them. Such sentiments -- for the half-credences of which I speak have never the full force of *thought* -- such sentiments are seldom thoroughly stifled unless by reference to the doctrine of chance, or, as it is technically termed, the Calculus of Probabilities. Now this Calculus is, in its essence, purely mathematical; and thus we have the anomaly of the most rigidly exact in science applied to the shadow and spirituality of the most intangible in speculation. (Egdar Allan Poe, The mistery of Marie Roget)

ψ,φ observables with expectations $\mu(\psi)$ and $\mu(\varphi)$

$(\sigma(\psi))^2 = [(\mu(\psi^2) - \mu(\psi)^2]$ **variance**

The **correlation coefficient** of ψ,φ is

$\rho(\psi,\varphi) = \text{covariance}(\psi,\varphi) / (\sigma(\psi)\,\sigma(\varphi))$
$= \mu\,[(\psi - \mu(\psi))(\varphi - \mu(\varphi))] / (\sigma(\psi)\,\sigma(\varphi))$
$= \mu\,[\psi\,\varphi - \mu(\psi)\mu(\varphi)] / (\sigma(\psi)\,\sigma(\varphi))$

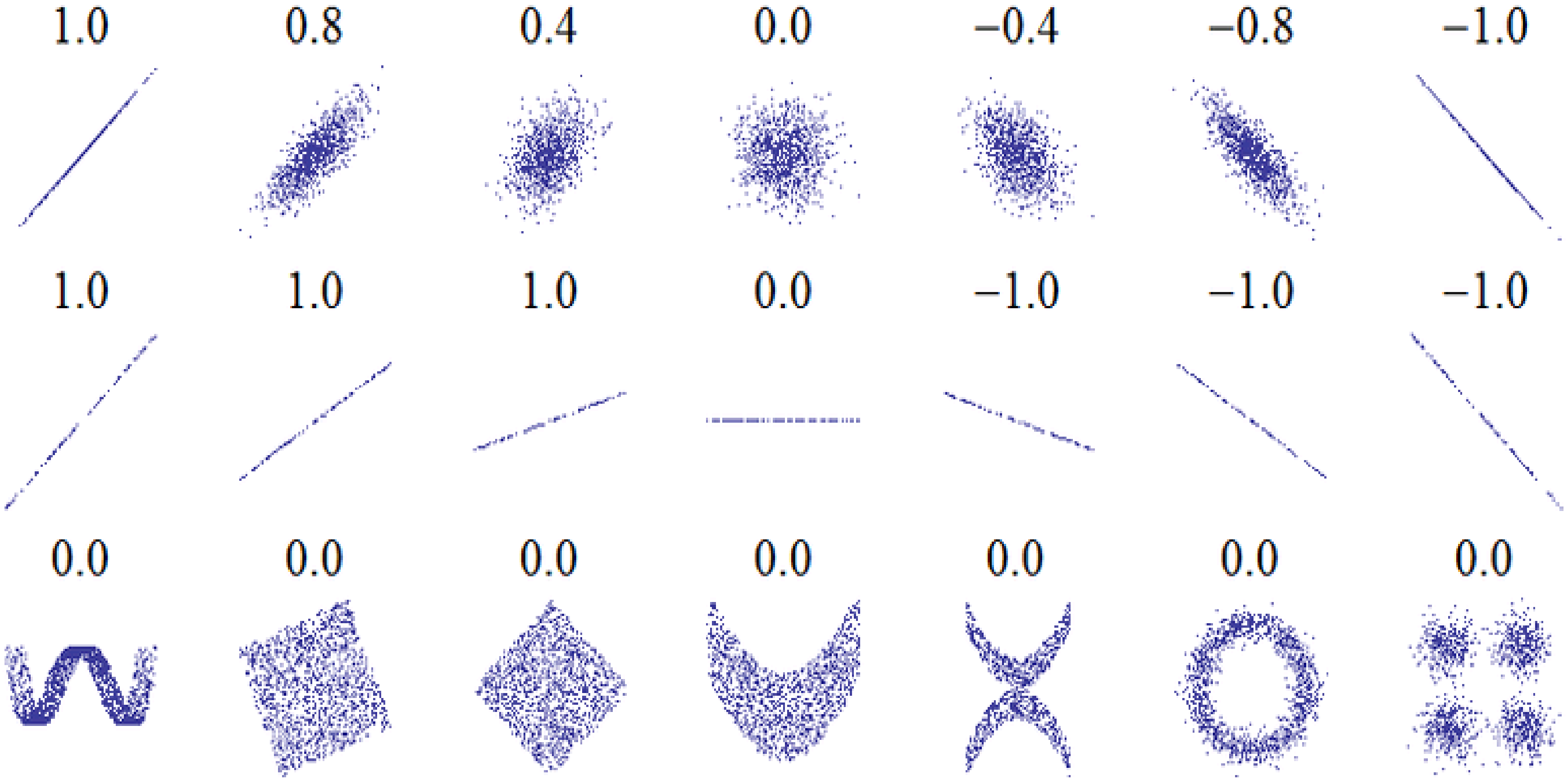
The correlation coefficient varies between -1 and 1 and equals 0 for independent variables but this is only a necessary condition (e.g. φ uniform on [-1,1] has zero correlation with its square)

If we have a series of n measurements of X and Y written as x(i) and y(i) where i = 1, 2, ..., n, then the Pearson product-moment correlation coefficient can be used to estimate the correlation of X and Y . The Pearson coefficient is also known as the "sample correlation coefficient". The Pearson correlation coefficient is then the best estimate of the correlation of X and Y . The Pearson correlation coefficient is written:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

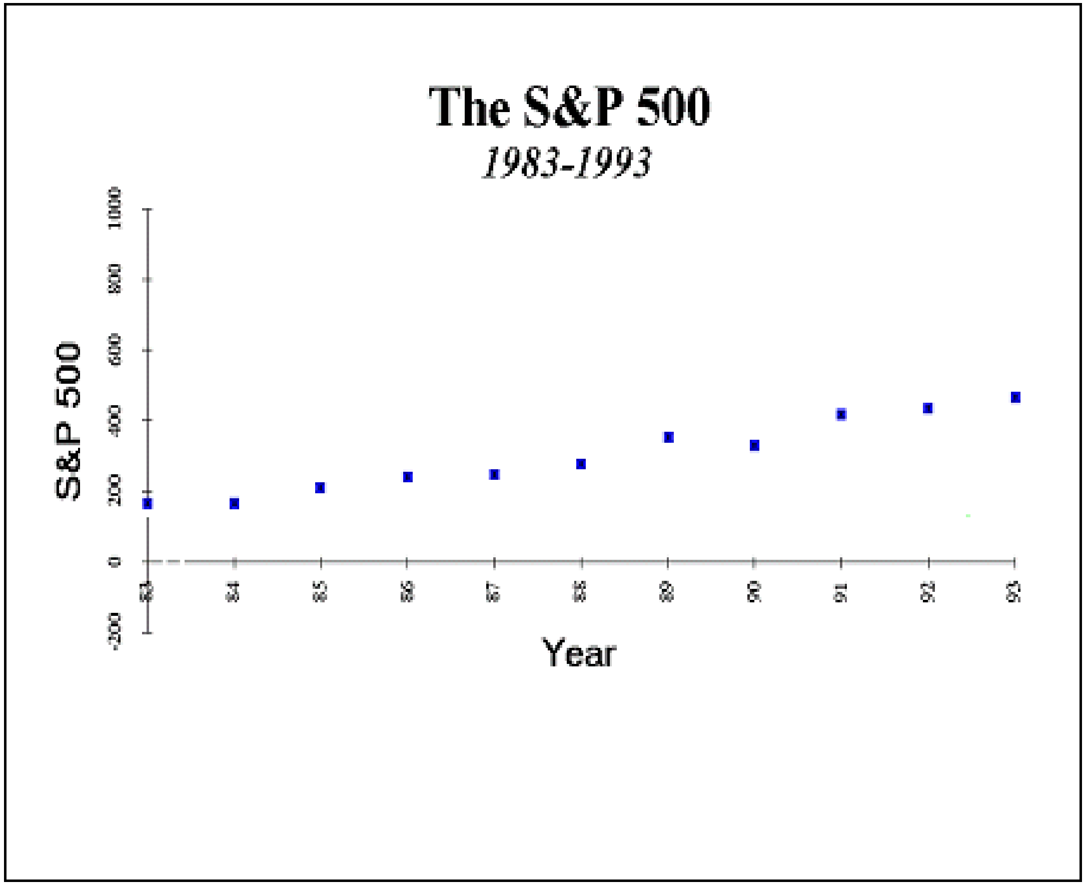
$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

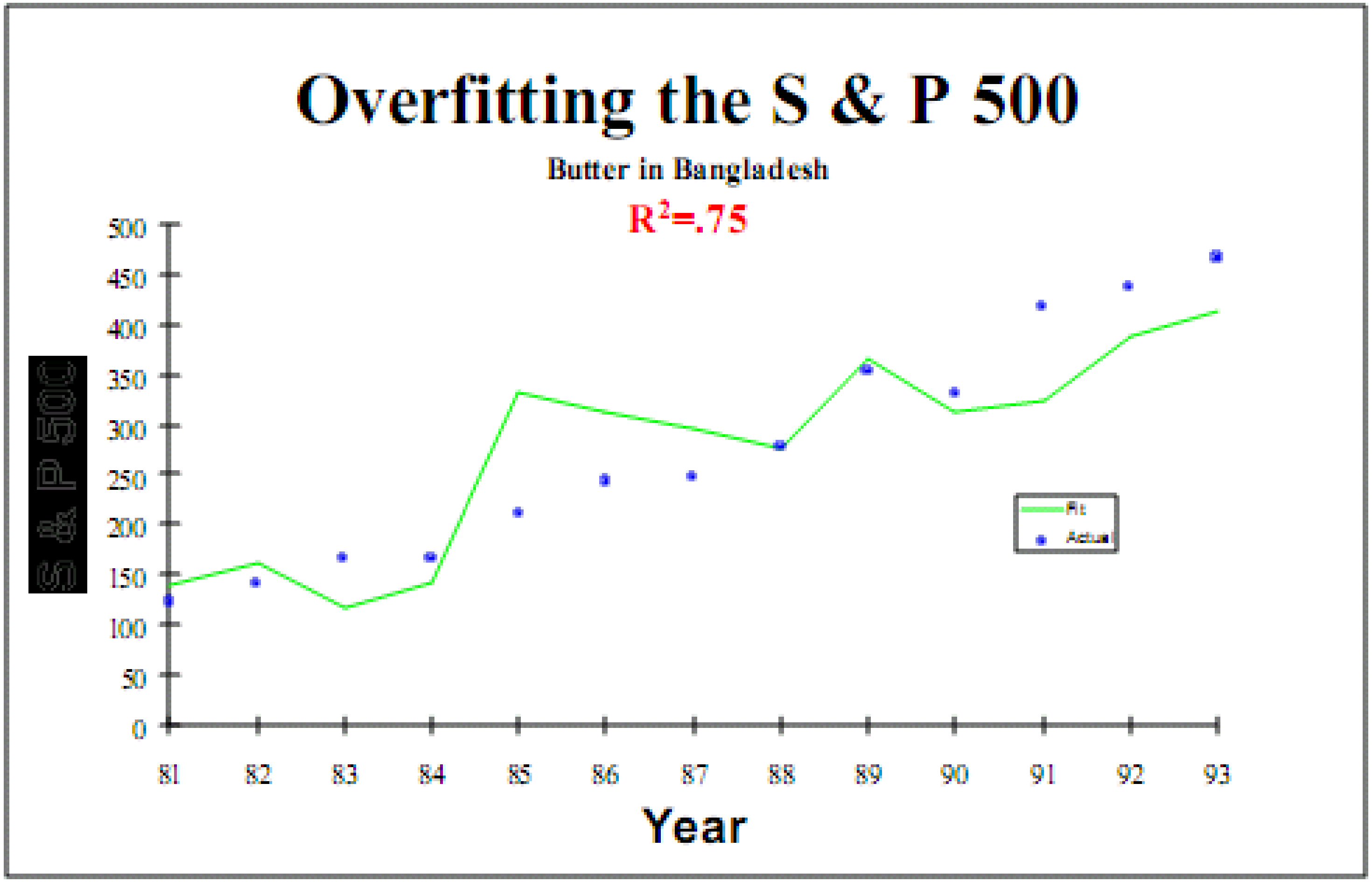
Correlation between two observables or series



Correlation and data-mining

We'll use the annual closing price of the S&P 500 index for the ten years from 1983 to 1993, shown in the chart below.



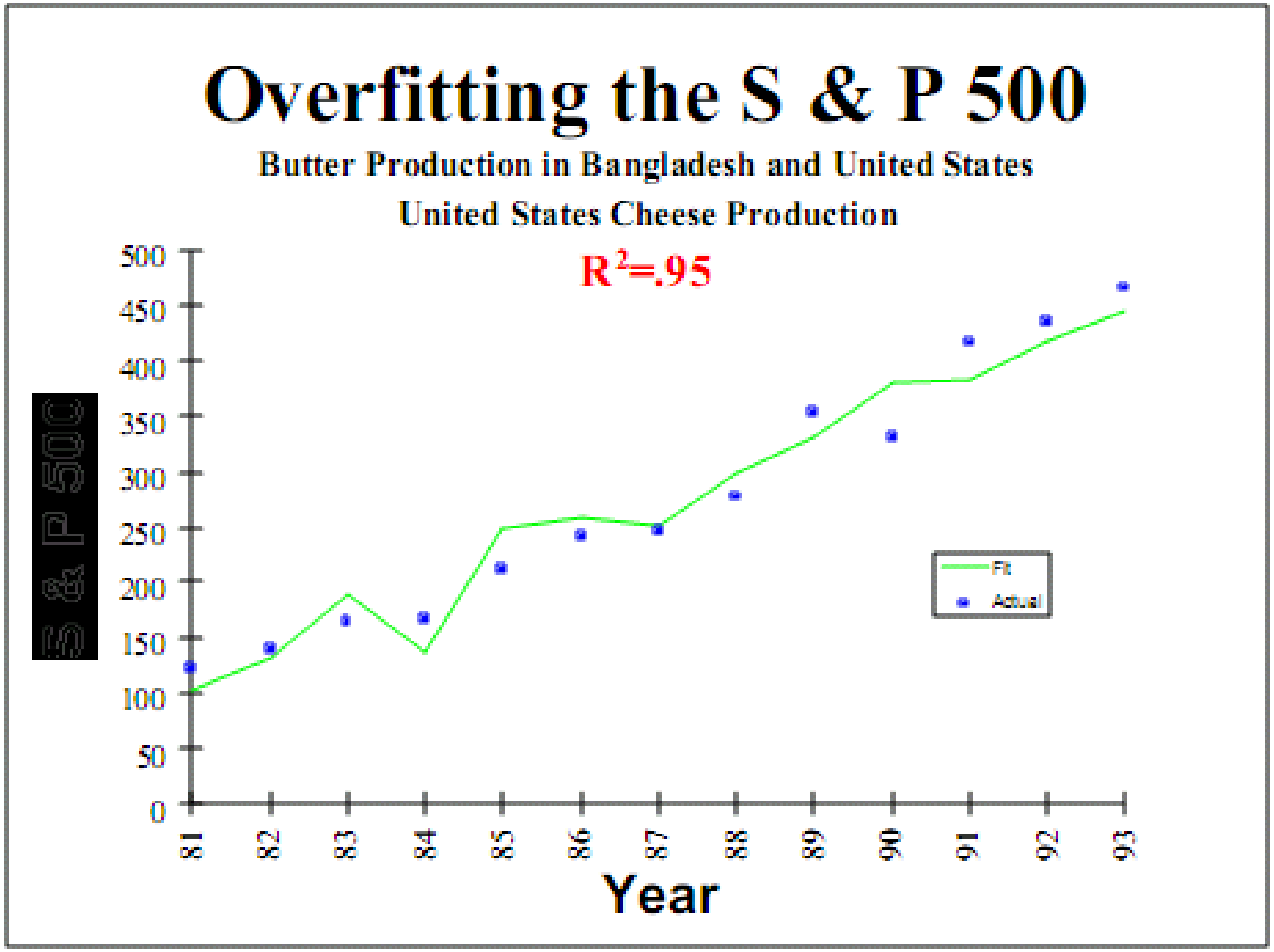


Stupid Data Miner Tricks: Overfitting the S&P 500

David J. Leinweber

THE JOURNAL OF INVESTING

Spring 2007

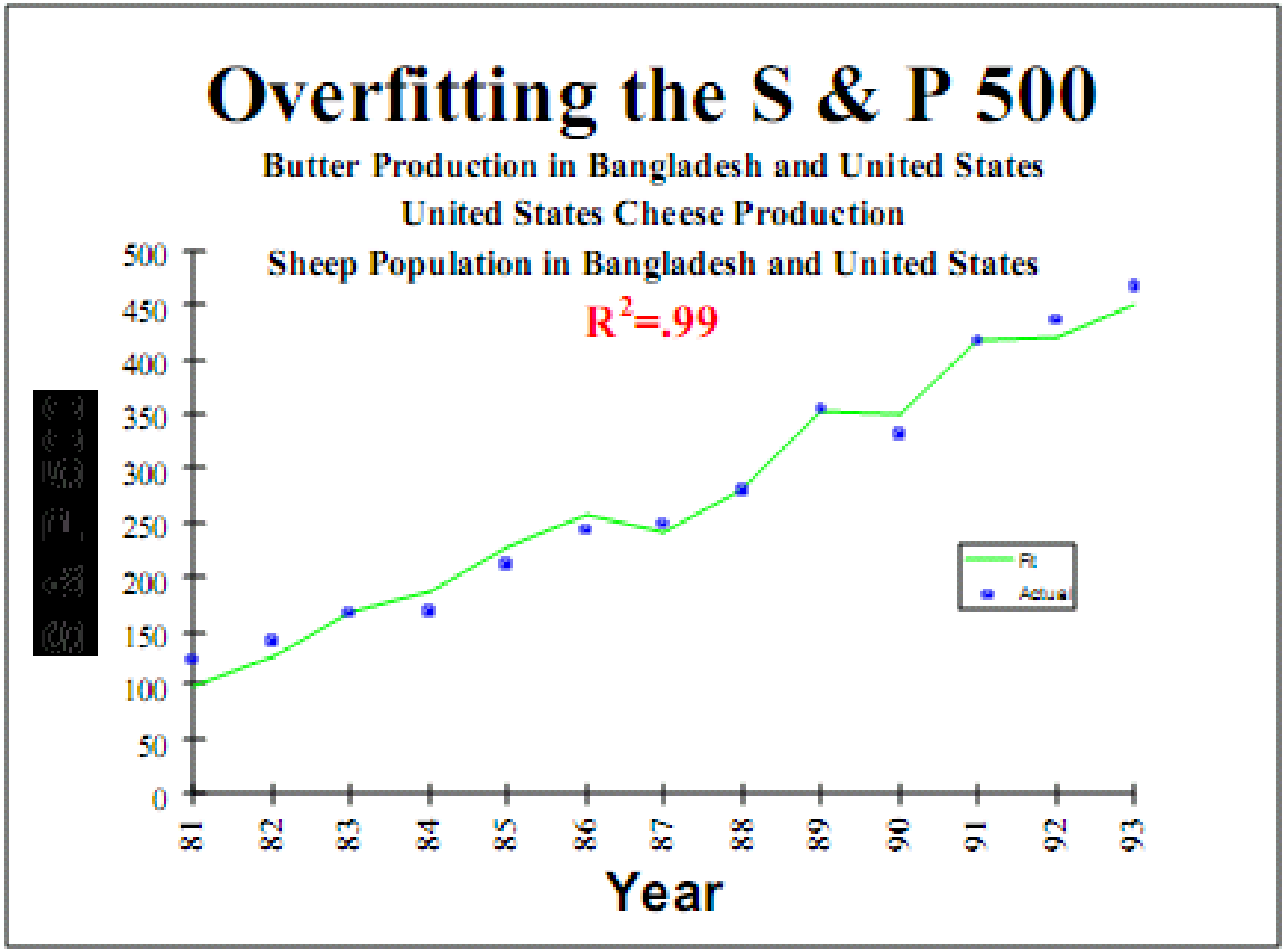


Stupid Data Miner Tricks: Overfitting the S&P 500

David J. Leinweber

THE JOURNAL OF INVESTING

Spring 2007



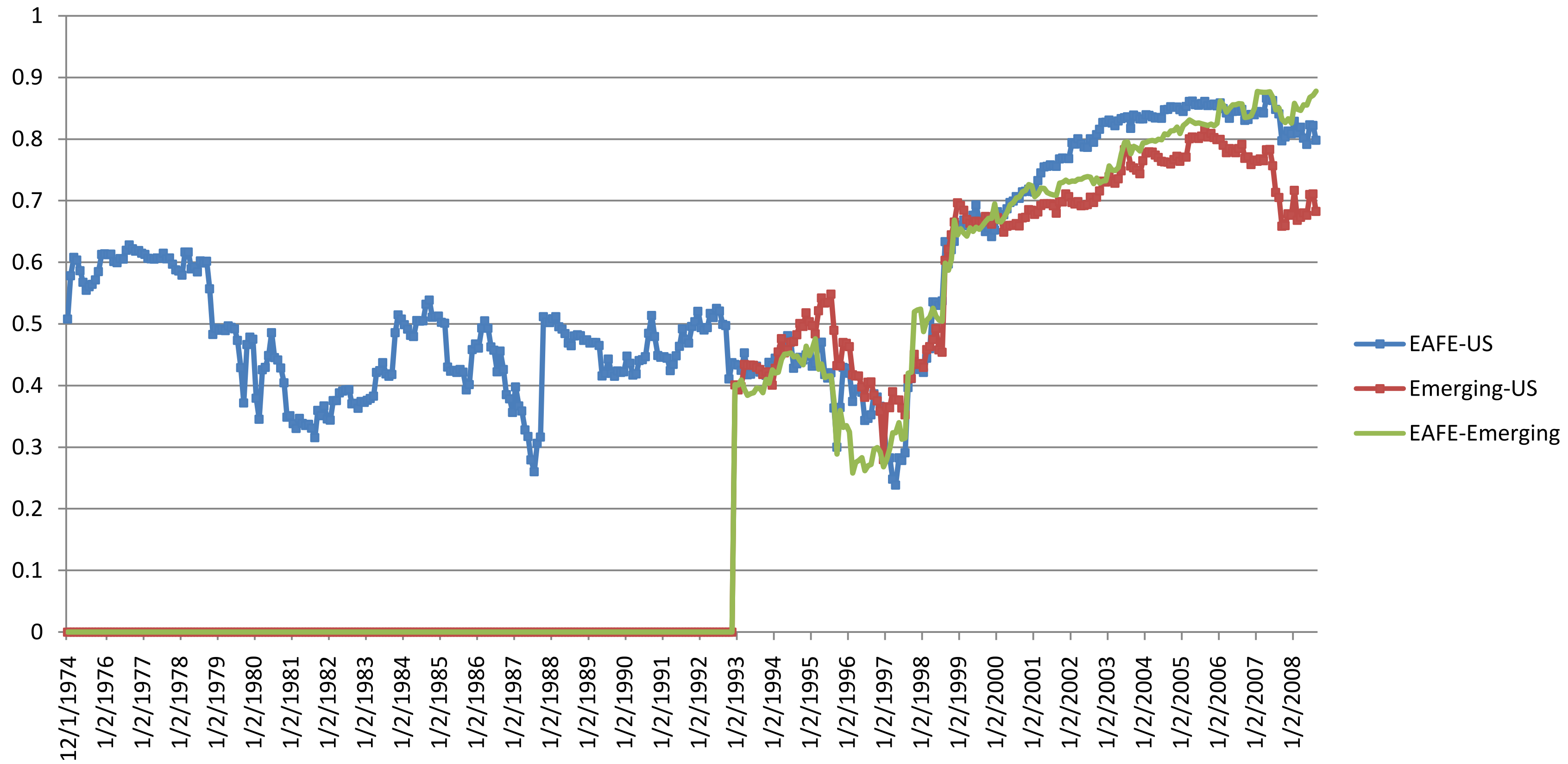
Stupid Data Miner Tricks: Overfitting the S&P 500

David J. Leinweber

THE JOURNAL OF INVESTING

Spring 2007

Historical correlation between stockmarkets



Correlation coefficients between rolling 5-year series of monthly returns of the indexes MSCI-Barra EAFE (Europe, Australasia, Far East), MSCI-U.S. and MSCI-Emerging Markets.

Mixing

Order n correlation coefficient:

$$c_n(\varphi,\psi) := \int \varphi.\psi\circ T^n d\mu - \int \varphi d\mu \int \psi d\mu$$

Ergodicity implies $\frac{1}{N}\sum_{0}^{N-1} c_n(\varphi,\psi) \longrightarrow 0$

Mixing requires that $c_N(\varphi,\psi) \longrightarrow 0$

namely φ and φ T^n become independent of each other as n→∞

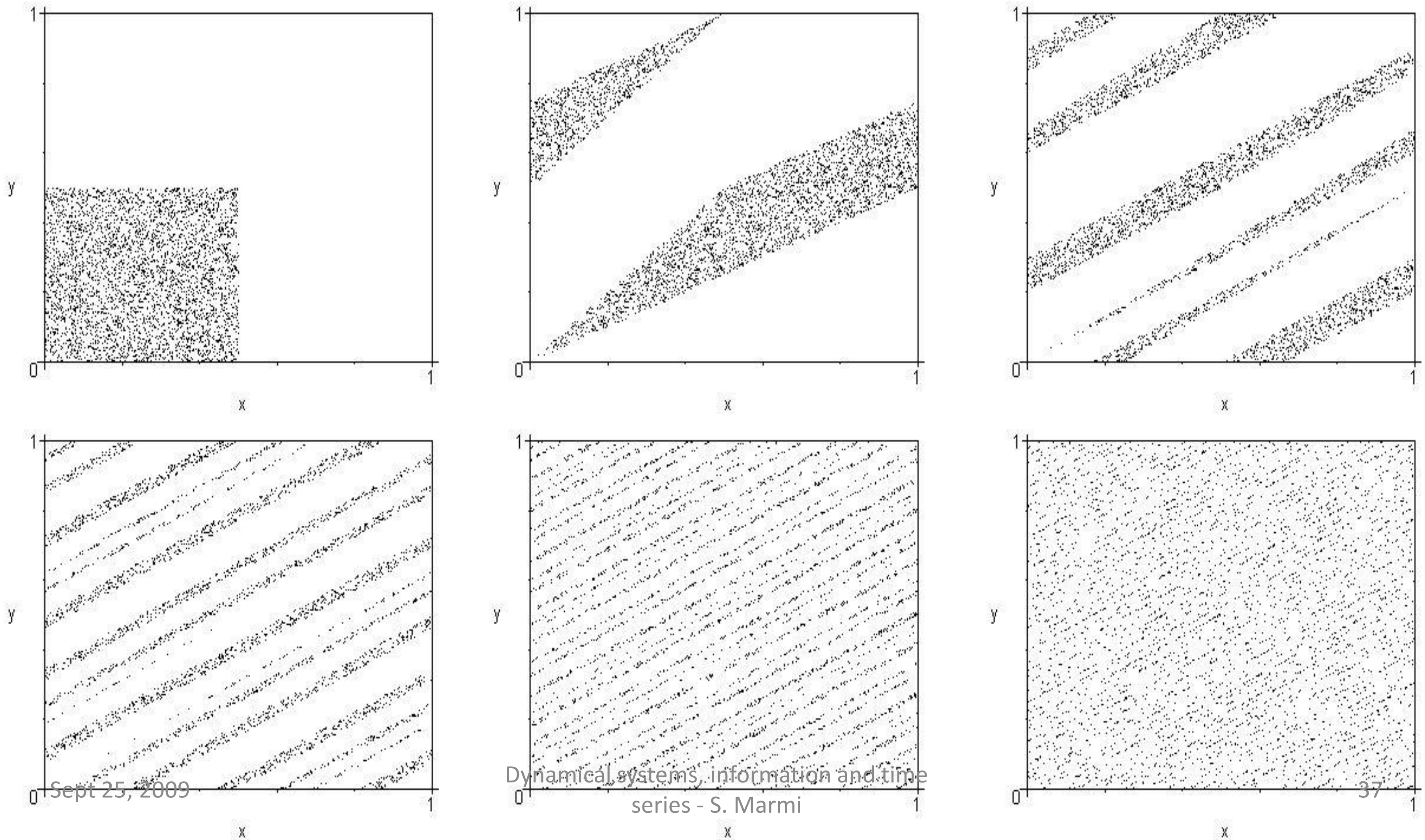
Strong vs. weak mixing

- *Strongly mixing* systems are such that for every E, F, we have $\mu(T^{-n}(E) \cap F) \rightarrow \mu(E)\mu(F)$ as n tends to infinity; the Bernoulli shift is a good example. Informally, this is saying that shifted sets become asymptotically independent of unshifted sets.

- *Weakly mixing* systems are such that for every E, F, we have $\mu(T^{-n}(E) \cap F) \rightarrow \mu(E)\mu(F)$ as n tends to infinity *after excluding a set of exceptional values of n of asymptotic density zero*.

- Ergodicity does not imply $\mu(T^{-n}(E) \cap F) \rightarrow \mu(E)\mu(F)$ but says that this is true for Cesaro averages: $n^{-1}\sum_{0\leq j\leq n-1} \mu(T^{-j}(E) \cap F) \rightarrow \mu(E)\mu(F)$

Mixing of hyperbolic automorphisms of the 2-torus (Arnold’s cat)



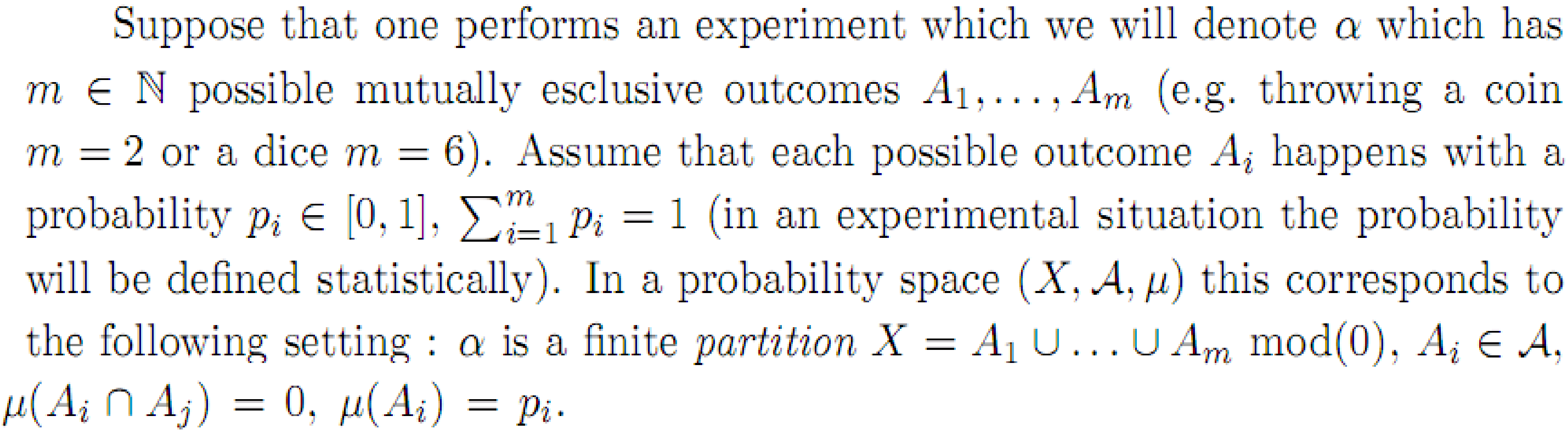
Entropy

In probability theory, *entropy* quantifies the uncertainty associated to a random process

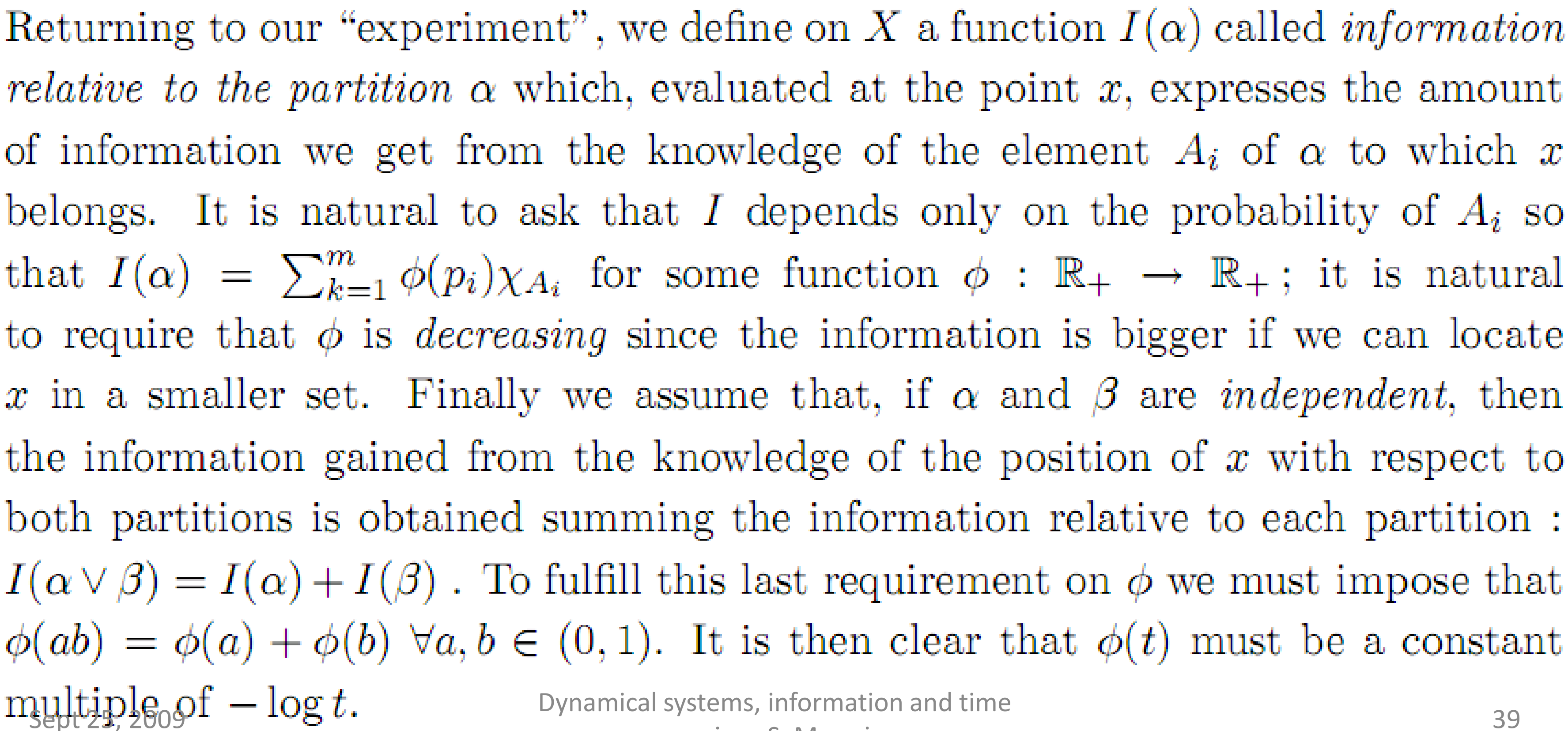
Consider an experiment with mutually esclusive outcomes $A=\{a_{1,} \ldots, a_k\}$

- Assume that the probability of a_i is p_i , $0\leq p_i \leq 1$, $p_1 + \ldots + p_k = 1$
- If a_1 has a probability very close to 1, then in most experiments the outcome would be a_1 thus the result is not very uncertain. One doea not gain much information from performing the experiment.
- One can quantify the "surprise" of the outcome as

  information= −log (probability)
- (the intensity of a perception is proportional to the logarithm of the intensity of the stimulus)



Suppose that one performs an experiment which we will denote α which has $m \in \mathbb{N}$ possible mutually esclusive outcomes $A_1, \ldots, A_m$ (e.g. throwing a coin $m = 2$ or a dice $m = 6$). Assume that each possible outcome A_i happens with a probability $p_i \in [0, 1]$, $\sum_{i=1}^{m} p_i = 1$ (in an experimental situation the probability will be defined statistically). In a probability space $(X, \mathcal{A}, \mu)$ this corresponds to the following setting : α is a finite *partition* $X = A_1 \cup \ldots \cup A_m \mod(0)$, $A_i \in \mathcal{A}$, $\mu(A_i \cap A_j) = 0$, $\mu(A_i) = p_i$.



Returning to our "experiment", we define on X a function $I(\alpha)$ called *information relative to the partition* α which, evaluated at the point x, expresses the amount of information we get from the knowledge of the element A_i of α to which x belongs. It is natural to ask that I depends only on the probability of A_i so that $I(\alpha) = \sum_{k=1}^{m} \phi(p_i)\chi_{A_i}$ for some function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$; it is natural to require that ϕ is *decreasing* since the information is bigger if we can locate x in a smaller set. Finally we assume that, if α and β are *independent*, then the information gained from the knowledge of the position of x with respect to both partitions is obtained summing the information relative to each partition : $I(\alpha \vee \beta) = I(\alpha) + I(\beta)$. To fulfill this last requirement on ϕ we must impose that $\phi(ab) = \phi(a) + \phi(b)$ $\forall a, b \in (0, 1)$. It is then clear that $\phi(t)$ must be a constant multiple of $-\log t$.

Entropy

The entropy associated to the experiment is

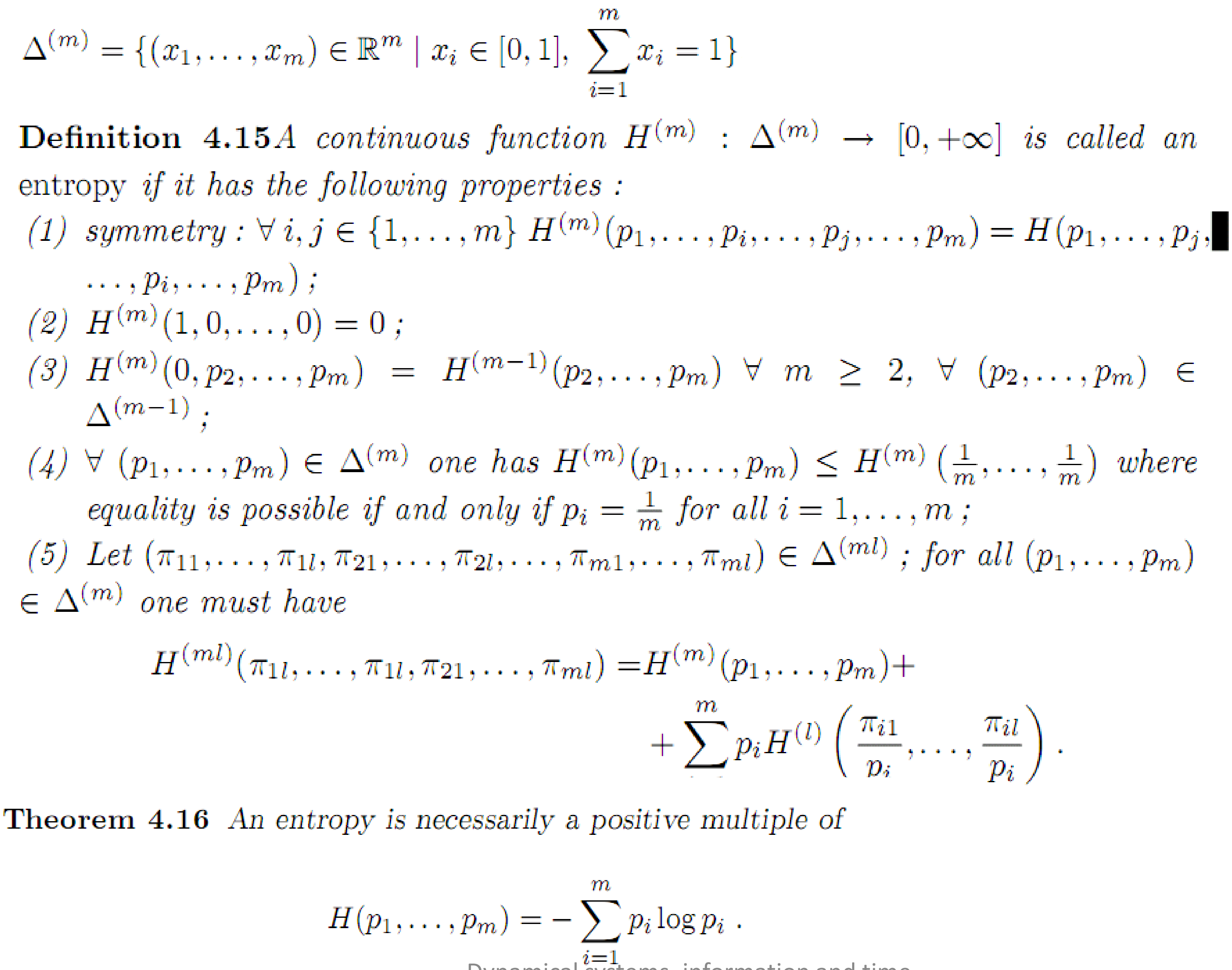
$H=-\sum p_i \log p_i$

Since

information = - Log (probability)

entropy is simply the expectation value of the information produced by the experiment

Uniqueness of entropy



$$\Delta^{(m)} = \{(x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_i \in [0,1],\ \sum_{i=1}^{m} x_i = 1\}$$

Definition 4.15 *A continuous function* $H^{(m)} : \Delta^{(m)} \to [0, +\infty]$ *is called an* entropy *if it has the following properties :*

(1) symmetry : $\forall\, i, j \in \{1, \ldots, m\}$ $H^{(m)}(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m) = H(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_m)$;

(2) $H^{(m)}(1, 0, \ldots, 0) = 0$;

(3) $H^{(m)}(0, p_2, \ldots, p_m) = H^{(m-1)}(p_2, \ldots, p_m)$ $\forall\ m \geq 2,\ \forall\ (p_2, \ldots, p_m) \in \Delta^{(m-1)}$;

(4) $\forall\ (p_1, \ldots, p_m) \in \Delta^{(m)}$ *one has* $H^{(m)}(p_1, \ldots, p_m) \leq H^{(m)}\left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$ *where equality is possible if and only if* $p_i = \frac{1}{m}$ *for all* $i = 1, \ldots, m$;

(5) Let $(\pi_{11}, \ldots, \pi_{1l}, \pi_{21}, \ldots, \pi_{2l}, \ldots, \pi_{m1}, \ldots, \pi_{ml}) \in \Delta^{(ml)}$; *for all* $(p_1, \ldots, p_m) \in \Delta^{(m)}$ *one must have*

$$H^{(ml)}(\pi_{1l}, \ldots, \pi_{1l}, \pi_{21}, \ldots, \pi_{ml}) = H^{(m)}(p_1, \ldots, p_m) + \\ + \sum^{m} p_i H^{(l)}\left(\frac{\pi_{i1}}{p_i}, \ldots, \frac{\pi_{il}}{p_i}\right).$$

Theorem 4.16 *An entropy is necessarily a positive multiple of*

$$H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i \ .$$

Entropy, coding and data compression

What does entropy measure?

Entropy quantifies the information content (namely the amount of randomness of a signal)

Entropy : a completely random binary sequence has entropy= $\log_2 2 = 1$ and cannot be compressed

Computer file= infinitely long binary sequence

Entropy = *best possible compression ratio*

Lempel-Ziv algorithm (Compression of individual sequences via variable rate coding, IEEE Trans. Inf. Th. 24 (1978) 530-536): does not assume knowledge of probability distribution of the source and achieves asymptotic compression ratio=entropy of source

The entropy of English

Is English is 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.

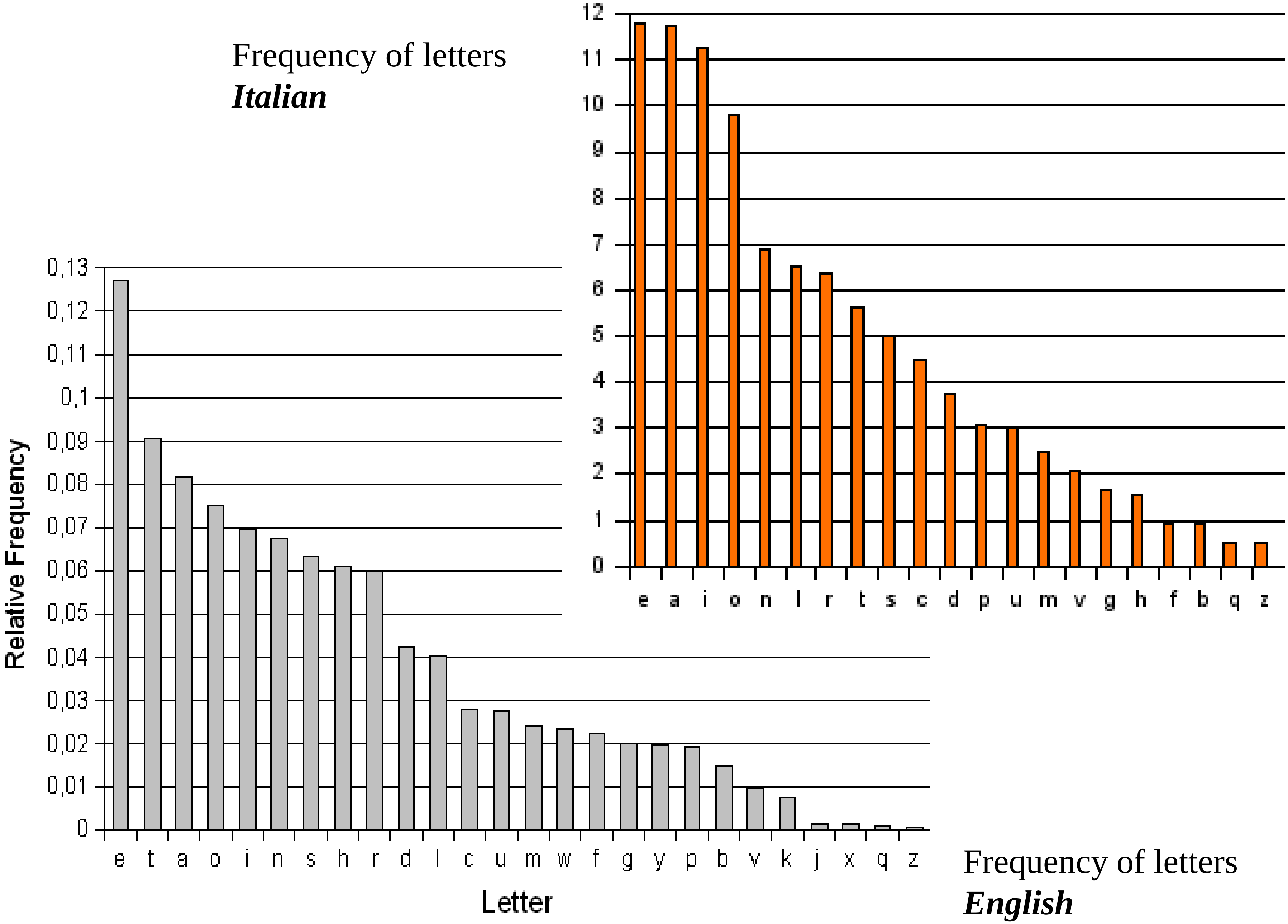
alphabet of English = 26 letters and the space symbol

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
Italian



Frequency of letters
English

From 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 $p(xi \mid xi-1, xi-2, xi-3)$ in correspondence of $27x27^3 = 531\,441$ entries for this table: 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