

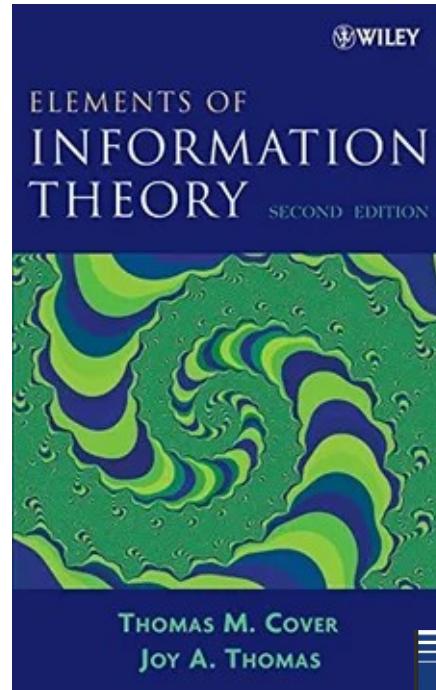
A tutorial on information theory

Juan M. López

Instituto de Física de Cantabria, CSIC and Universidad de
Cantabria, Spain

Outline

- 1.- What is information?
- 2.- How do we measure information?
- 3.- Information redundancy/correlations/patterns
- 4.- Information of processes
- 5.- Examples
- 6.- Maximum entropy principle (MaxEnt)



What is information?

J K I Q P J Q Y H T K N J H D I A E U K I Q P J Q Y H T K N J H D I A E U K I Q P J Q Y H T
J B L N L E B W I G R D Z U S V B K J B L N L E B W I G R D Z U S V B K J B L N L E B W I G
E S A F K S N D X V H O N G K O N G E S A F K S N D X V H O N G K O N G E S A F K S N D X V
Z Z N Y T A T H E N S N E T L M I J C Z N Y T A T H E N S N E T L M I J C Z N Y T A T H E N S
A M D X R N Z A B K U Y W O A E S V A M D X R N Z A B K U Y W O A E S V A M D X R N Z A B K U
R G H A I F M L O J C R D K I S T U R G H A I F M L O J C R D K I S T U R G H A I F M L O J C
B A D M O R Q X S Y D N E Y V G A P B A D M O R Q X S Y D N E Y V G A P B A D M O R Q X S Y D
E L E S N A U T Z M I G L O S A N G E L E S N A U T Z M I G L O S A N G E L E S N A U T Z M
N B Q T L N P G E D F A H L U R B W N B Q T L N P G E D F A H L U R B W N B Q T L N P G E D
F J K E Z C L D X H N P I M H D U P F J K E Z C L D X H N P I M H D U P F J K E Z C L D X H N P
A H V R J I W A B C B R U S S E L S A H V R J I W A B C B R U S S E L S A H V R J I W A B C
E C B D M S N K H D J I Q A F R Y G E C B D M S N K H D J I Q A F R Y G E C B D M S N K H D
I N F A P C A R A C A S V M J X J B I N F A P C A R A C A S V M J X J B I N F A P C A R A C
H O L M U O X T S L Z G E S T O C K H O L M U O X T S L Z G E S T O C K H O L M U O X T S L Z
P H G T X V C Q S U B E R L I N Z Q P H G T X V C Q S U B E R L I N Z Q P H G T X V C Q S U B
E D I N R M A K R G H F K O V J R L E D I N R M A K R G H F K O V J R L E D I N R M A K R G H F
I W S U O N S Y M L J D A F E C A B I W S U O N S Y M L J D A F E C A B I W S U O N S Y M L
C H K A M P A R I S G B B X L M D F C H K A M P A R I S G B B X L M D F C H K A M P A R I S G
A P O R E S B O J X F E C H S I N G A P O R E S B O J X F E C H S I N G A P O R E S B O J X F
P J Z N L Q L M L P N U D K T Q C Z P J Z N L Q L M L P N U D K T Q C Z P J Z N L Q L M L P N
E A F C R B A R C E L O N A D J U B E A F C R B A R C E L O N A D J U B E A F C R B A R C E L
T M I X S Z N S A K R V M R G R K U T M I X S Z N S A K R V M R G R K U T M I X S Z N S A K R V
D R K D V R C O I M Z D E H N E W Y O R K D V R C O I M Z D E H N E W Y O R K D V R C O I M Z
W E A B J X A U R J B L Y A C X D S W E A B J X A U R J B L Y A C X D S W E A B J X A U R J B
N C M E H F P M O S C O W R Q F V H N C M E H F P M O S C O W R Q F V H N C M E H F P M O S C
U K I Q P J Q Y H T K N J H D I A E U K I Q P J Q Y H T K N J H D I A E U K I Q P J Q Y H T K N J H D
D B L N L E B W I G R D Z U S V B K J B L N L E B W I G R D Z U S V B K J B L N L E B W I G R D Z
E S A F K S N D X V H O N G K O N G E S A F K S N D X V H O N G K O N G E S A F K S N D X V H
Z Z N Y T A T H E N S N E T L M I J C Z N Y T A T H E N S N E T L M I J C Z N Y T A T H E N S N E

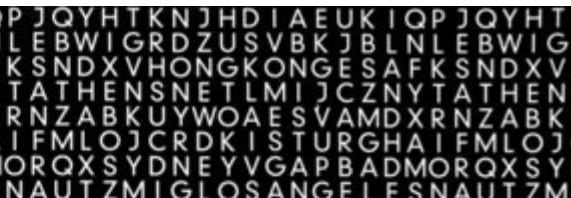
What is information?

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.



This block contains two examples of blind text. The first example is a single-line sequence of letters and numbers: P J Q Y H T K N J H D I A E U K I Q P J Q Y H T L E B W I G R D Z U S V B K J B L N L E B W I G K S N D X V H O N G K O N G E S A F K S N D X V I T A T H E N S N E T L M I J C Z N Y T A T H E N S R N Z A B K U Y W O A E S V A M D X R N Z A B K U I F M L O J C R D K I S T U R G H A I F M L O J C O R Q X S Y D N E Y V G A P B A D M O R Q X S Y D N A U T Z M I G L O S A N G E L E S N A U T Z M L N P G Z C L D J I W A M S N K P C A R I U O X T X V C G R M A K O N S Y M P A R E S B C



This block contains two examples of blind text. The second example is a multi-line sequence of letters and numbers:

```
J Z N L Q L M P N O D K I Q C Z P J Z N L Q L M
E A F C R B A R C E L O N A D J U B E A F C R B A R
T M I X S Z N S A K R V M R G R K U T M I X S Z N S
D R K D V R C O I M Z D E H N E W Y O R K D V R C C
W E A B J X A U R J B L Y A C X D S W E A B J X A U
N C M E H F P M O S C O W R Q F V H N C M E H F P M
U K I Q P J Q Y H T K N J H D I A E U K I Q P J Q Y
D B L N L E B W I G R D Z U S V B K J B L N L E B W
E S A F K S N D X V H O N G K O N G E S A F K S N D
Z N Y T A T H E N S N E T L M I J C Z N Y T A T H
```

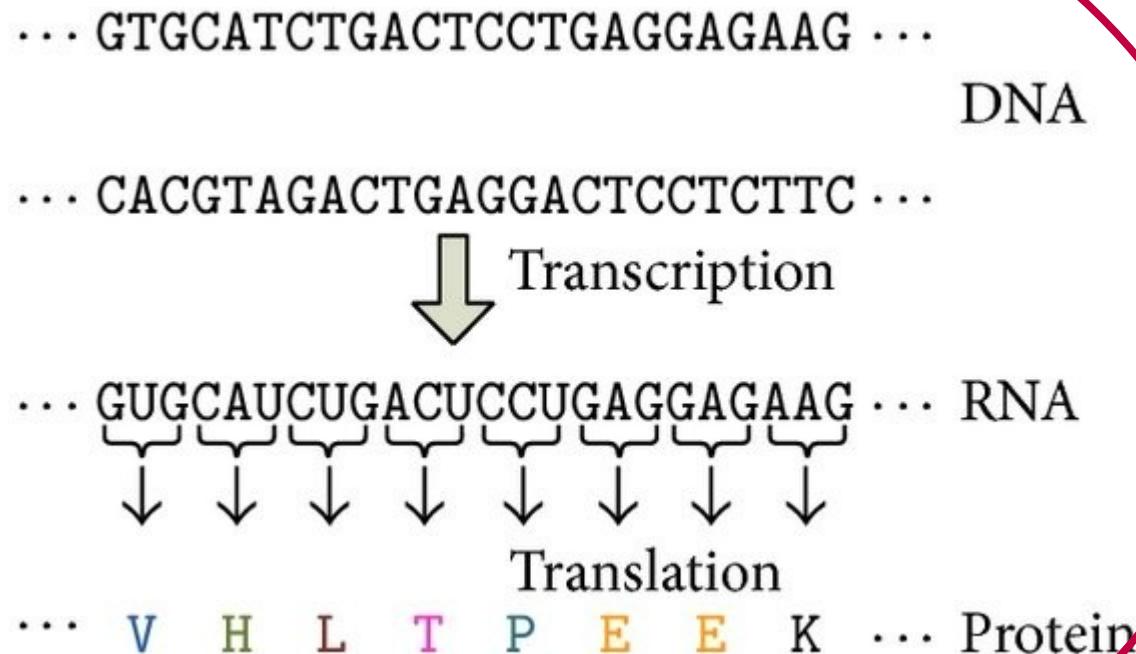
Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

What is information?

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schrift abgelesen und man kann prüfen, ob es Sinnvolles oder Nonsense ist und was es bedeutet.

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schrift abgelesen und man kann prüfen, ob es Sinnvolles oder Nonsense ist und was es bedeutet.



meaning. This text should show this place. If you read this text, is there no information? Is there mere nonsense like "Huardest gefbt like this gives you information" or are written and an impression of all letters of the alphabet and a language. There is no need for words should match the language.

meaning. This text should show this place. If you read this text, is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

What is information?

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird dor

Grauwert der Schrift

kann man prüfen, lesen ist und wie s

Dies ist ein Blindt

vieles über

Example of avoiding complex expressions

Grauwert

kann man prüfen,

lesen ist und wie s

... GTGCATCTGACTCCTGAGGAGAAG ...

DNA

uld show
this text,
!? Is there
rdest gef-
formation
impression
abet and
o need for
age.

ould show
this text,
here a
urn"?
ut the
look.
uld be
ntent,

```
def square_numbers(numbers):
    """Return a list of the squares of the numbers."""
    squares = []
    for number in numbers:
        squares.append(number**2)
    return squares

def main():
    numbers = [1, 2, 3, 4]
    result = square_numbers(numbers)
    print(result)
```

What is information?

3.141592653589793238462643
3832795028841971693993751
0582097494459230781640628
6208998628034825342117067
9821480865132823066470938
4460955058223172535940812
8481117450284102701938521
1055596446229489549303819
6442881097566593344612847

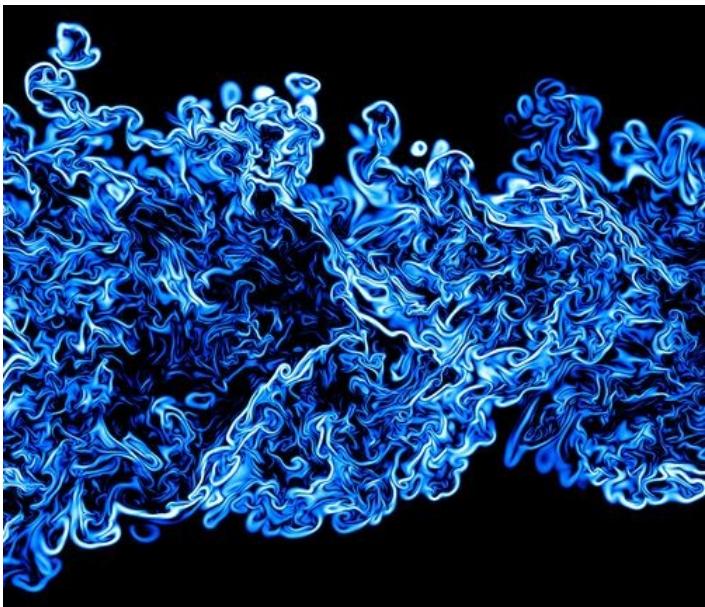
7052066222233333588433225803349333333333:
0744994453367566433003308809337233754428:
6597084493309954033063306274332733979606:
2604802473397862733403357540337033333334:
2968978243304044633207337893394733969674:
8060569023322459733582033933025033452269:
3333333498033333350995683337048033333333:

1110111011000110110100010110111101100
101001101110011001100010111101101111000
0011000100000101010111010011011011010101
101001011111101111111000110110101001000
000100011110110100011000110111111101000
00101110011111110110001011010101110010100
010010001010110111010000100111101000001
0011111100010101011100000100111001111000
0110010011110001111010001101100001011100
1111001100101010001110000001010011111000
101000101100001100101111101100101110011
110101101101111111100000001100001111000
1010011001101001011010011100111010100101
0011101001010011111110000001011010101110
0101111111011011000110110101100010110101
0100001110100110111100110011000101111011
10110100001100010000010101011101001101101
0010100010100101111101111111100011010101
0010010100010001111011010001100011011111

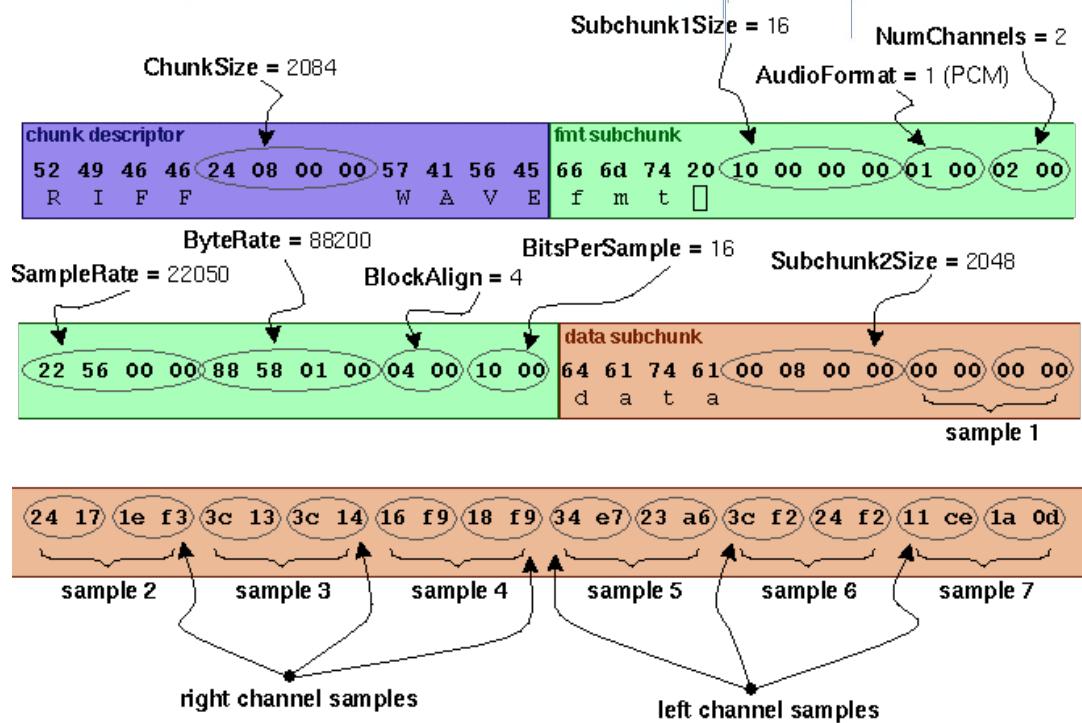
What is information?



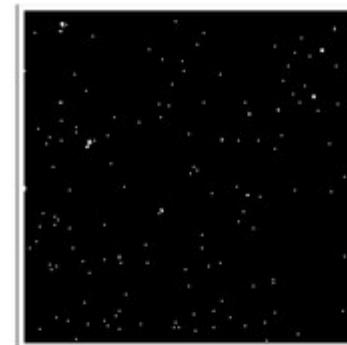
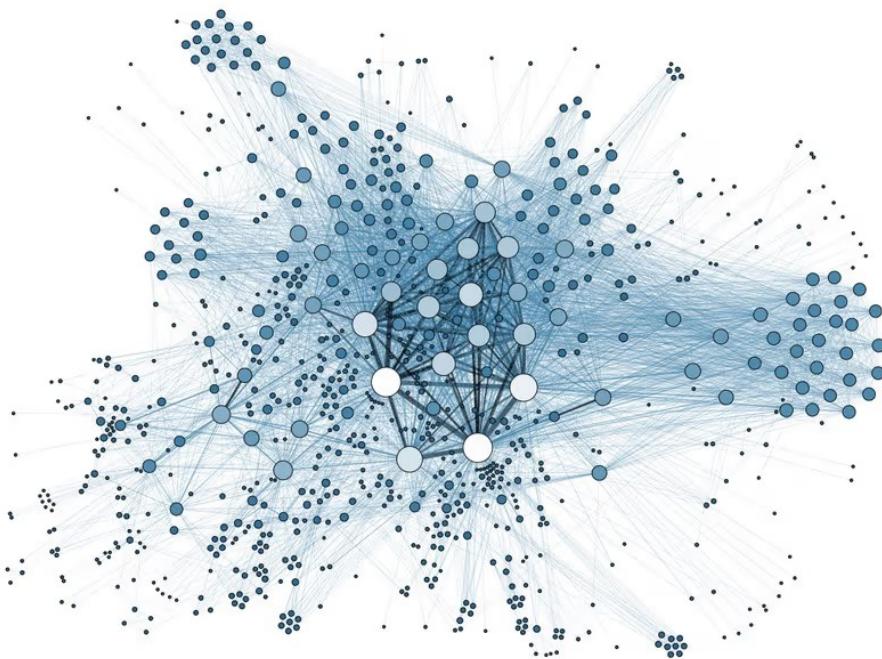
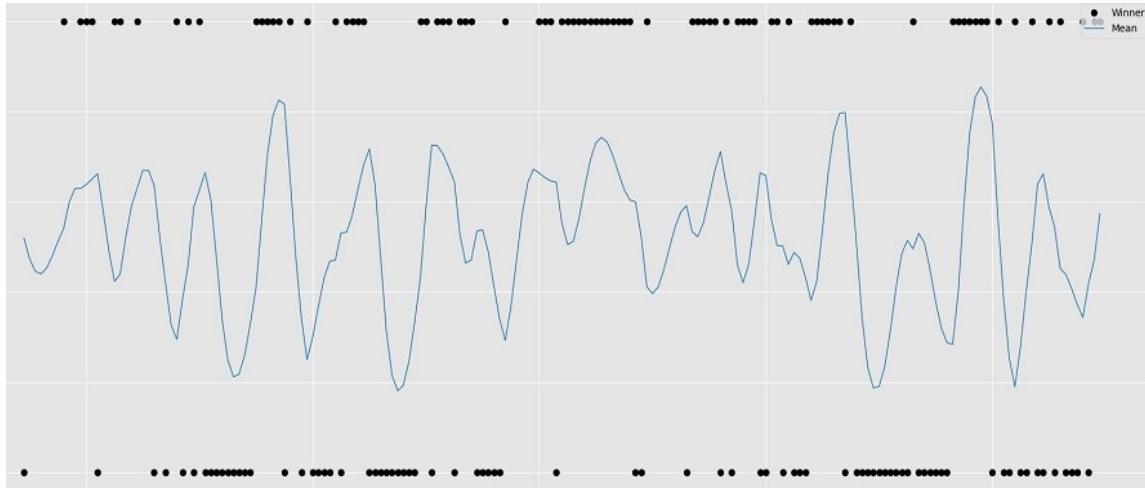
What is information?



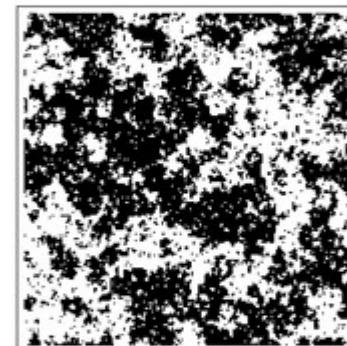
What is information?



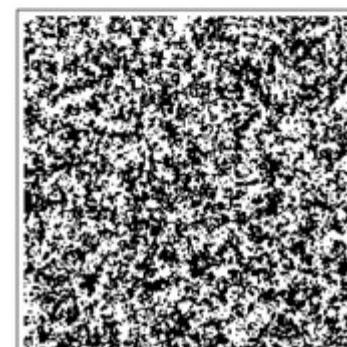
What is information?



$T < T_c$



$T \sim T_c$

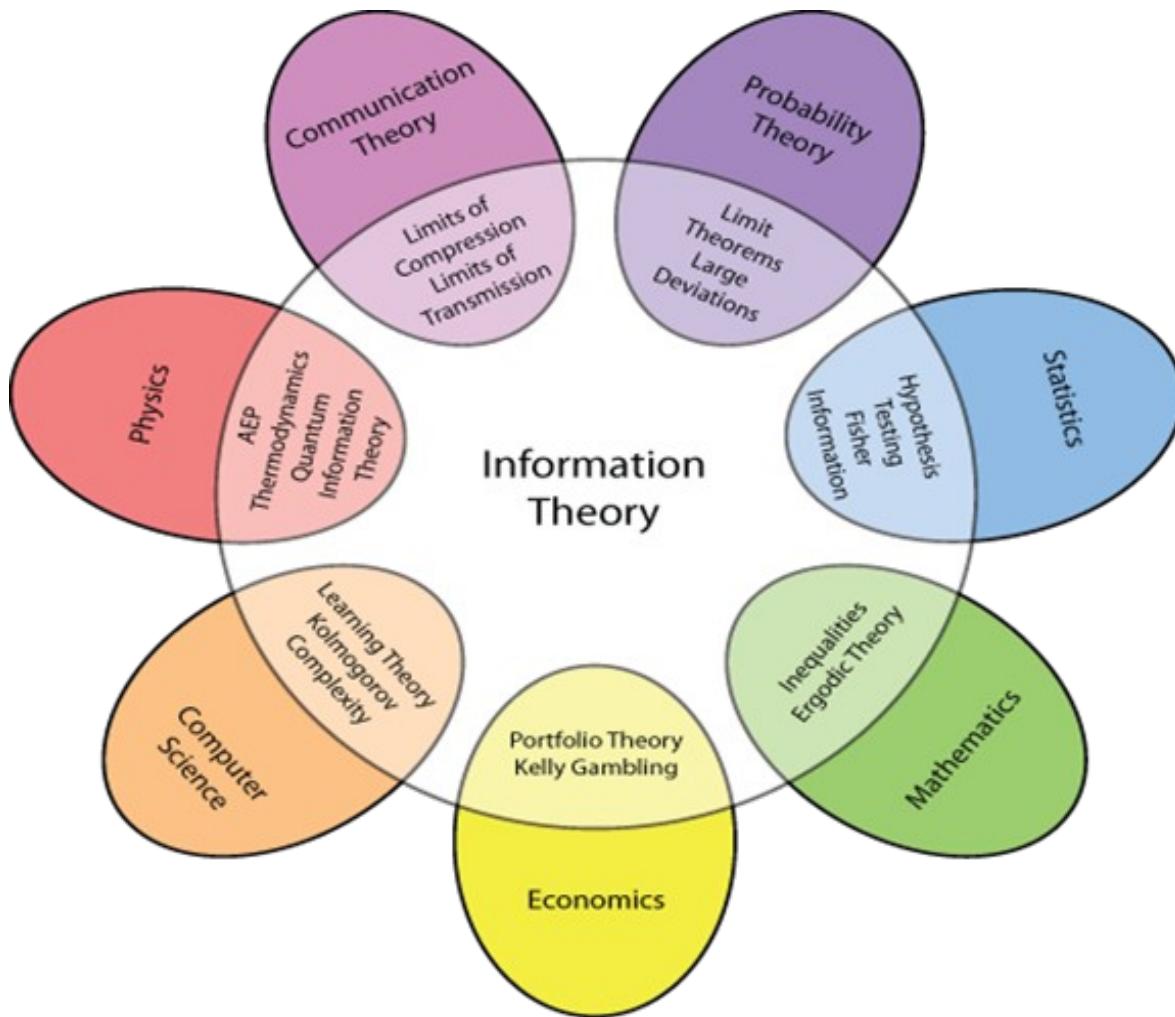


$T > T_c$

What is information theory?

1. The mathematical theory of **communication, compression, complexity, correlation, uncertainty, information processing, ...**
 - What does communication mean? What are the fundamental limits on communication? How can codes for communication be derived and compared? How can data be represented in a compressed way? How much information does one distribution provide about another?
2. The mathematical theory underlying **much of statistics and machine learning**
 - The most general, "right" metrics for quantifying degrees of uncertainty, divergence, and correlation among distributions
3. A nexus among fields...

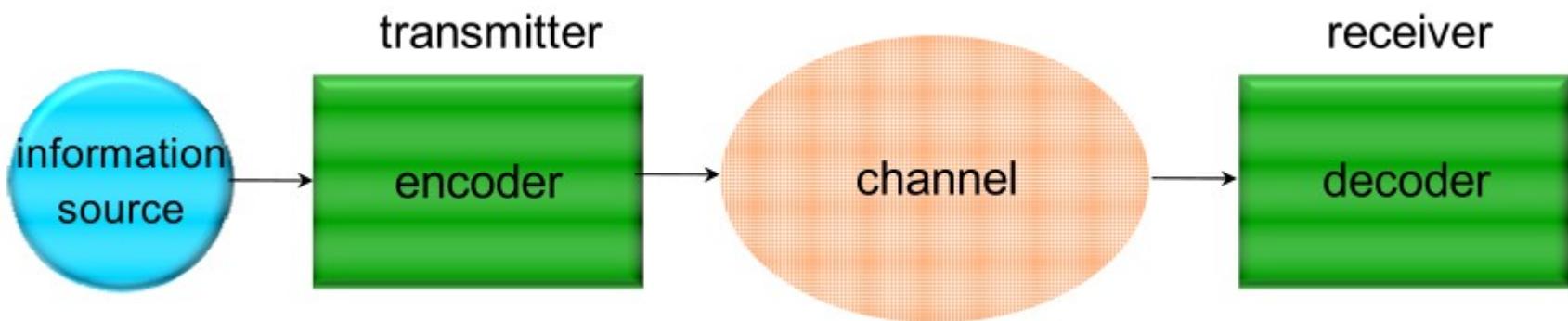
What is information?



What is information?

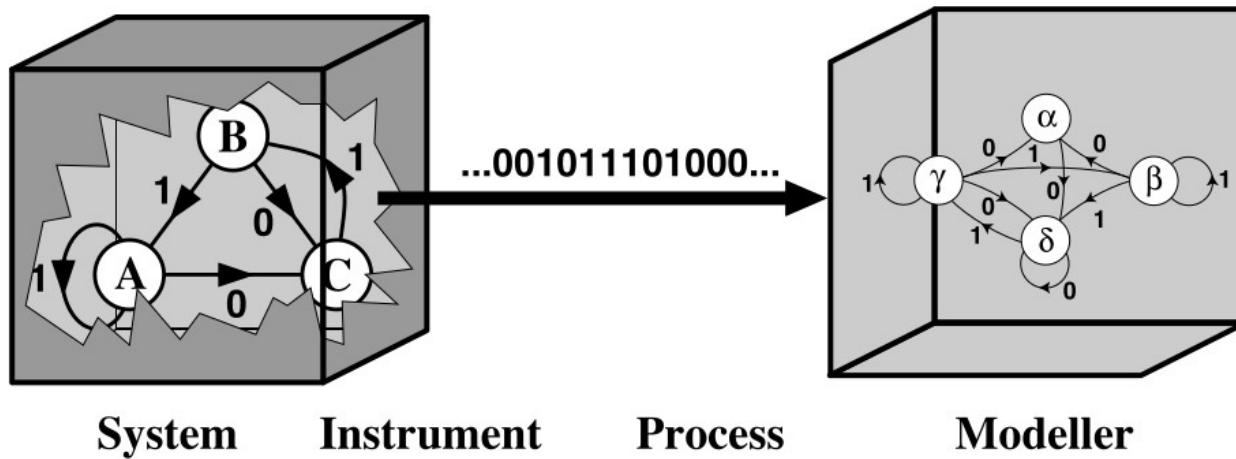
Shannon theory of information

“The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.”



Claude Shannon, *A Mathematical Theory of Communication* (The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948)

What is information?



The Learning Channel

What is information?

How much information?

Alice (A) is in a state of ignorance about a certain variable X which is known to Bob (B). She anticipates that the answer $X \in \mathcal{X}$ can be one of $n = |\mathcal{X}|$ possible ones. \mathcal{X} is called the *alphabet* and X is a *symbol*

One way to quantify the information content of X is to count the number of binary questions (yes/no) that A needs to pose to B in order to know the answer X

Therefore, the number N_Q of binary questions needed to dispel A's ignorance is an operative definition of the information content of X , and it is measured in bits

What is information?

Take for example the case $\mathcal{X} = \{a, b, c, d\}$. Then A may ask a first question

Q_1 : is $X \in \{a, b\}$ or not?

and depending on the answer, A may ask

Q_2 : if $X \in \{a, b\}$ is $X = a$ or not?

Else, if $X \notin \{a, b\}$ is $X = c$ or not?

The answers to these two questions reveal the correct outcome X . Hence the information is $N_Q = 2$ bits. Yet there are many other ways in which A could ask questions, and hence N_Q could vary accordingly.

What is information?

Take for example the case $\mathcal{X} = \{a, b, c, d\}$. Then A may ask a first question

Q_1 : is $X \in \{a, b\}$ or not?

and depending on the answer, A may ask

Q_2 : if $X \in \{a, b\}$ is $X = a$ or not?

Else, if $X \notin \{a, b\}$ is $X = c$ or not?

The answers to these two questions reveal the correct outcome X . Hence the information is $N_Q = 2$ bits. Yet there are many other ways in which A could ask questions, and hence N_Q could vary accordingly.

What is information?

Q'_1 : is $X = a$ or not?

only if $X \neq a$ A will need to pose a further question. Then she may ask:

Q'_2 : is $X = b$ or not?

Only if the result is no, she will need to ask

Q'_3 : is $X = c$ or not?

in which case the number of binary questions can be $N_Q(a) = 1$, $N_Q(b) = 2$ or $N_Q(c) = N_Q(d) = 3$, depending on the value of X . Indeed, $N_Q(X)$ is a random variable, because it is a function of X .

Therefore, it makes sense to define a measure of information content as the expected number of binary questions that are needed to elicit the value of X .

$$\mathbb{E}[N_Q] = \sum_{x \in \chi} p_x N_Q(x)$$

What is information?

Blackboard calculations

What is information?

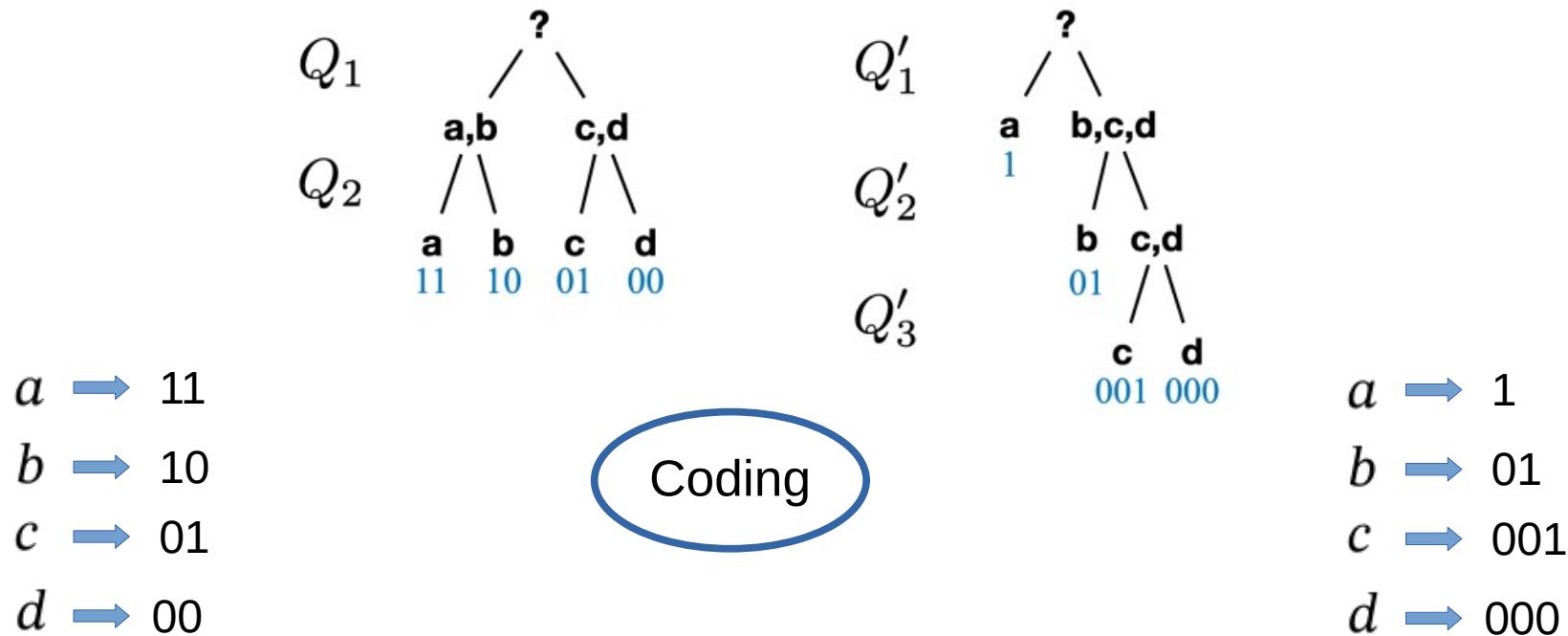
The optimal way of answering questions is different in the two cases. The minimal expected number of binary questions that A needs to pose to elicit X is a measure of her irreducible ignorance about X . Hence, we provisionally define

The information content $H[X]$ of a random variable X is the *minimal* expected number of binary questions needed to elicit the value of X ,

$$H[X] = \min_Q \mathbb{E}[N_Q] \quad (16.2)$$

where the expected value is taken with respect to the distribution $P\{X = x\} = p_x$ that defines the state of knowledge on X , and the minimum is taken over all possible ways of posing yes/no questions.

What is information?



Notice that each codeword has length $\ell_Q(X) = N_Q(X)$ which is equal to the number of binary questions needed to elicit X under protocol Q .

Therefore the problem of finding the code that is *expected* to use the least number of bits (i.e. that minimises $\mathbb{E} [\ell_Q]$) is exactly the same as the problem of finding the best way to pose questions. The fact that these two apparently different problems — A posing questions to B optimally and B transmitting answers to A efficiently — have the same solution, is interesting.

What is information?

The minimal number of binary questions needed to elicit X , or equivalently the expected length of the optimal code for X , is given by the Shannon entropy

$$H[X] = \mathbb{E} [\log_2 1/p_X] = - \sum_{x \in \mathcal{X}} p_x \log_2 p_x$$

What is information?

Asymptotic Equipartition Property

Let $\underline{X} = (X_1, \dots, X_n)$ be independent draws from a discrete distribution $p(x)$ ($X_i \in \chi$ with $|\chi| < +\infty$). Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p(\underline{X}) = -H[X] \quad (14.4)$$

in probability.

$$H[X] = - \sum_{x \in \chi} p(x) \log p(x)$$

What is information?

Asymptotic Equipartition Property

For any $\epsilon > 0$, one can define the set of ϵ -typical sequences as

$$A_n^\epsilon = \left\{ \underline{X} : \left| \frac{1}{n} \log p(\underline{X}) + H[X] \right| < \epsilon \right\}$$

Then an equivalent way to state the AEP is that

1. By definition, all ϵ -typical sequences are equally likely: $P(\underline{X}) \sim e^{-nH[X]}$ for all $\underline{X} \in A_n^\epsilon$
2. As a consequence of the law of large numbers, a random sequence is almost surely an ϵ -typical sequence

$$P\{A_n^\epsilon\} > 1 - \epsilon.$$

3. As a consequence, the number of ϵ -typical sequences is

$$|A_n^\epsilon| \sim e^{nH[X]}.$$

What is information?

Asymptotic Equipartition Property

Therefore, the number of typical samples $|A_n^\epsilon| \sim e^{nH[X]}$ is much smaller than the number of all possible samples, which is $|\chi|^n = e^{n\log|\chi|}$, whenever the distribution differs from the uniform one $p(x) = 1/|\chi|$, for which one has $H[X] = \log|\chi|$. To put it differently, the probability that any sequence \underline{X} is typical is exponentially small in n .

What is information?

Example: Binary random variable X (Biased Coin)

$$\mathcal{X} = \{0, 1\} \quad \Pr(1) = p \text{ & } \Pr(0) = 1 - p$$

$$H(X) ?$$

Binary entropy function:

$$H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$$

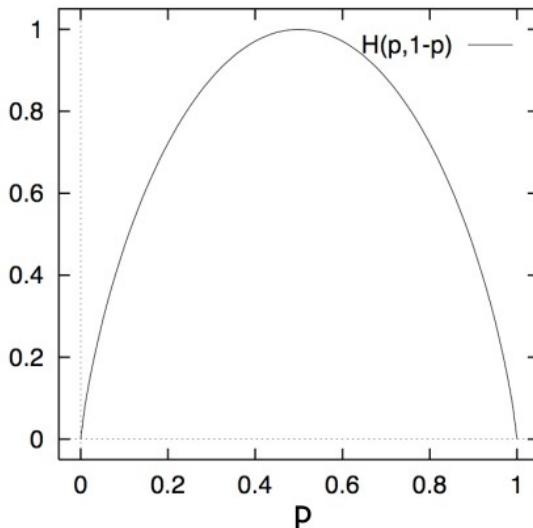
Fair coin: $p = \frac{1}{2}$

$$H(p) = 1 \text{ bit}$$

Completely biased coin: $p = 0$ (or 1)

$$H(p) = 0 \text{ bits}$$

Recall: $0 \cdot \log 0 = 0$



What is information?

Example: IID Process over four events

Entropy: $H(X) = \frac{7}{4}$ bits

At each stage, ask questions that are most informative.

Choose partitions of event space that give “most random”
measurements.

Theorem:

Entropy gives the smallest number of questions
to identify an event, on average.

What is information?

Interpretations of Shannon Entropy:

Observer's degree of *surprise* in outcome of a random variable

Uncertainty *in* random variable

Information required to *describe* random variable

A measure of *flatness* of a distribution

What is information?

Two random variables: $(X, Y) \sim p(x, y)$

Joint Entropy: Average uncertainty in X and Y occurring

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y)$$

Independent:

$$X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$$

What is information?

Conditional Entropy: Average uncertainty in X , knowing Y

$$H(X|Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x|y)$$

$$H(X|Y) = H(X, Y) - H(Y)$$

Not symmetric: $H(X|Y) \neq H(Y|X)$

What is information?

Common Information Between Two Random Variables:

$$X \sim p(x) \text{ & } Y \sim p(y)$$

$$(X, Y) \sim p(x, y)$$

Mutual Information:

$$I(X; Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

What is information?

Properties:

- (1) $I(X; Y) \geq 0$
- (2) $I(X; Y) = I(Y; X)$
- (3) $I(X; Y) = H(X) - H(X|Y)$
- (4) $I(X; Y) = H(X) + H(Y) - H(X, Y)$
- (5) $I(X; X) = H(X)$
- (6) $X \perp Y \Rightarrow I(X; Y) = 0$

Interpretations:

- Information one variable has about another
- Information shared between two variables
- Measure of dependence between two variables

What is information?

Properties:

- (1) $I(X; Y) \geq 0$
- (2) $I(X; Y) = I(Y; X)$
- (3) $I(X; Y) = H(X) - H(X|Y)$
- ★ (4) $I(X; Y) = H(X) + H(Y) - H(X, Y)$
- (5) $I(X; X) = H(X)$
- (6) $X \perp Y \Rightarrow I(X; Y) = 0$

Interpretations:

Information one variable has about another

Information shared between two variables

Measure of dependence between two variables

What is information?

Why is mutual information better than correlation?

<https://docs.google.com/file/d/1bx6uPNuFkxVc8Fc1C5hVN3hZoaq1fw/preview>

What is information?

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x . How much this impacts on the efficiency of the questions she's going to ask?

What is information?

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x . How much this impacts on the efficiency of the questions she's going to ask?

Given q , A is going to effectively encode B's answers in such a way that answer x will require $\log_2 1/q_x$ bits, so the number of questions she will ask, on average, is

$$\mathbb{E} \left[\log_2 \frac{1}{q} \right] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{1}{q_x}$$

What is information?

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x . How much this impacts on the efficiency of the questions she's going to ask?

Given q , A is going to effectively encode B's answers in such a way that answer x will require $\log_2 1/q_x$ bits, so the number of questions she will ask, on average, is

$$\mathbb{E} \left[\log_2 \frac{1}{q} \right] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{1}{q_x}$$

the difference between this and the most efficient way of asking questions, which requires $\mathcal{H}[p]$ bits, is

$$D_{KL}[p\|q] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{p_x}{q_x}$$

which is known as the Kullback-Leibler divergence or relative entropy

What is information?

Data processing system



Nature



Camera

CD

What is information?

Basic Result: Data Processing Inequality

- Let X and Y be random variables and let g be any function. We have:

$$\text{MI}(g(X); Y) \leq \text{MI}(X; Y)$$



What?????????????????????????????

What is information?

Basic Result: Data Processing Inequality

- Let X and Y be random variables and let g be any function. We have:

$$\text{MI}(g(X); Y) \leq \text{MI}(X; Y)$$

- What does this tell us about the meaning of "information"?
- MI quantifies information in X about Y , regardless of whether that information is "usable" or "accessible" to some particular agent.
 - Coming up with new MI measures which account for accessibility of information is a hot topic in information theory!

What is information?

Some facts about mutual information

1. Accounts for any type of co-relation

- Statistical correlation ~ linear only
- Information measures nonlinear correlation

2. Broadly applicable:

- Many systems don't have "energy", physical modeling precluded
- Information defined: social, biological, engineering, ... systems

3. Comparable units across different systems:

- Correlation: Meters v. volts v. dollars v. ergs v. ...
- Information: bits.

4. Probability theory ~ Statistics ~ Information

5. Complex systems:

- Emergent patterns!
- We don't know these ahead of time

Information redundancy/patterns

Emma Woodh*use, hands*m^e, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her. *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e***e**, **h**h** ***l***n***i***l***s***r***o***a***o***e***i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W***o***s***i***l***a***g***n***t***a***e***v***

Information redundancy/patterns

Information redundancy



Information correlation/statistical dependence along
data/patterns



Dynamical model/Inference/Function

Information redundancy/patterns

Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her. *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*riage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**, **h**h** **l**n**i**l**s**r**o**a**o**e**i**
a***c***n***S***e***y***s***d***s***a***r***e***n***
W***o***s***i***l***a***g***n***t***a***e***v***

Information redundancy/patterns

Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her. *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffection*ate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*riage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e***e**, **h***h** ***l**n**i**l**s**r**o**a**o**e**i**
a***c***n***S***e***y***s***d***s***a***r***e***n***
W***o***s***i***l***a***g***n***t***a***e***v***

Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty one years in the world with very little to distress or vex her. She was the youngest of the two daughters of a most affectionate, indulgent father; and had, in consequence of her sister's marriage, been mistress of his house from a very early period. Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses; and her place had been supplied by an excellent woman as governess, who had fallen little short of a mother in affection. Sixteen years had Miss Taylor been in Mr Woodhouse's family, less as a governess than a friend, very

Example: Letter Frequencies

i	a_i	p_i	
1	a	0.0575	a
2	b	0.0128	b
3	c	0.0263	c
4	d	0.0285	d
5	e	0.0913	e
6	f	0.0173	f
7	g	0.0133	g
8	h	0.0313	h
9	i	0.0599	i
10	j	0.0006	j
11	k	0.0084	k
12	l	0.0335	l
13	m	0.0235	m
14	n	0.0596	n
15	o	0.0689	o
16	p	0.0192	p
17	q	0.0008	q
18	r	0.0508	r
19	s	0.0567	s
20	t	0.0706	t
21	u	0.0334	u
22	v	0.0069	v
23	w	0.0119	w
24	x	0.0073	x
25	y	0.0164	y
26	z	0.0007	z
27	-	0.1928	-

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

[Book by David MacKay](#)

Example: Letter Frequencies

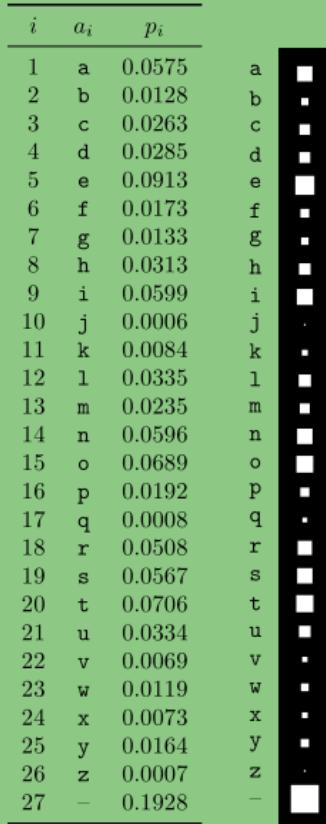


Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

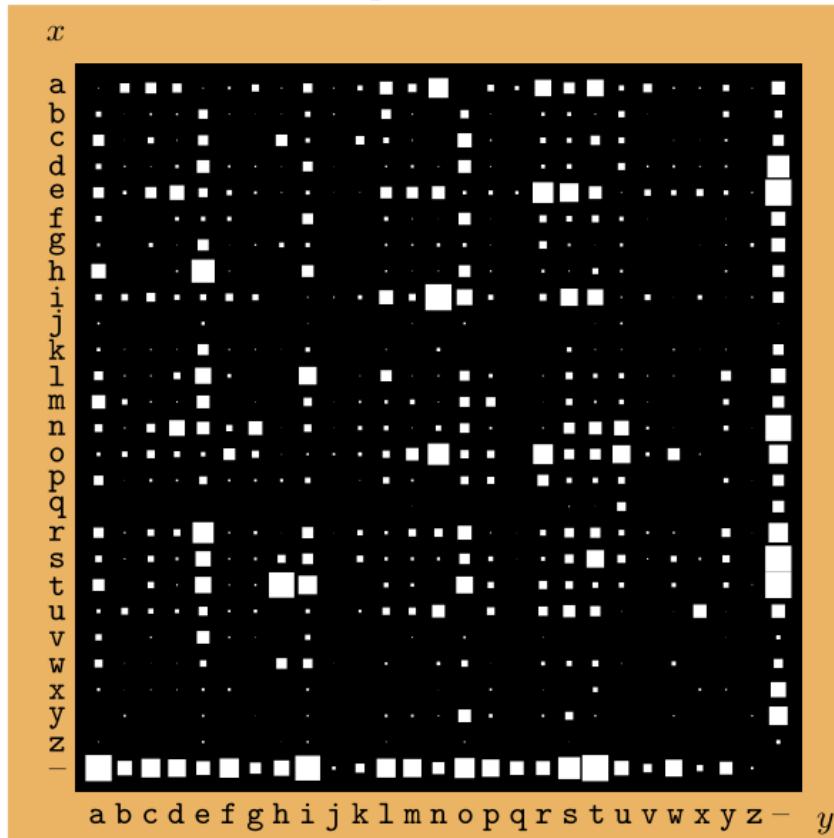


Figure 2.2. The probability distribution over the 27×27 possible bigrams xy in an English language document, *The Frequently Asked Questions Manual for Linux*.

[Book by David MacKay](#)

Information in processes

Block Entropy:

$$H(L) = H(\Pr(s^L)) = - \sum_{s^L \in \mathcal{A}} \Pr(s^L) \log_2 \Pr(s^L)$$

Monotonic increasing: $H(L) \geq H(L - 1)$

Adding a random variable cannot decrease entropy:

$$H(S_1, S_2, \dots, S_L) \leq H(S_1, S_2, \dots, S_L, S_{L+1})$$

No measurements, no information: $H(0) = 0$

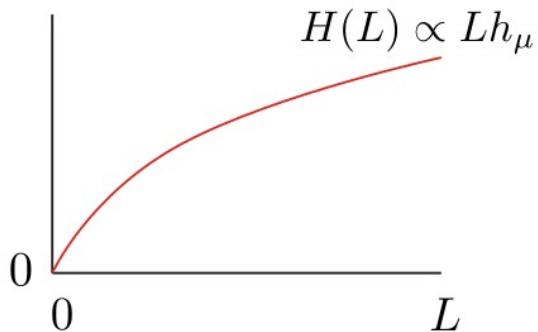
Information in processes

Entropy Rates for Stationary Stochastic Processes:

Entropy per symbol is given by the **Source Entropy Rate**:

$$h_\mu = \lim_{L \rightarrow \infty} \frac{H(L)}{L}$$

(When limits exists.)



Interpretations:

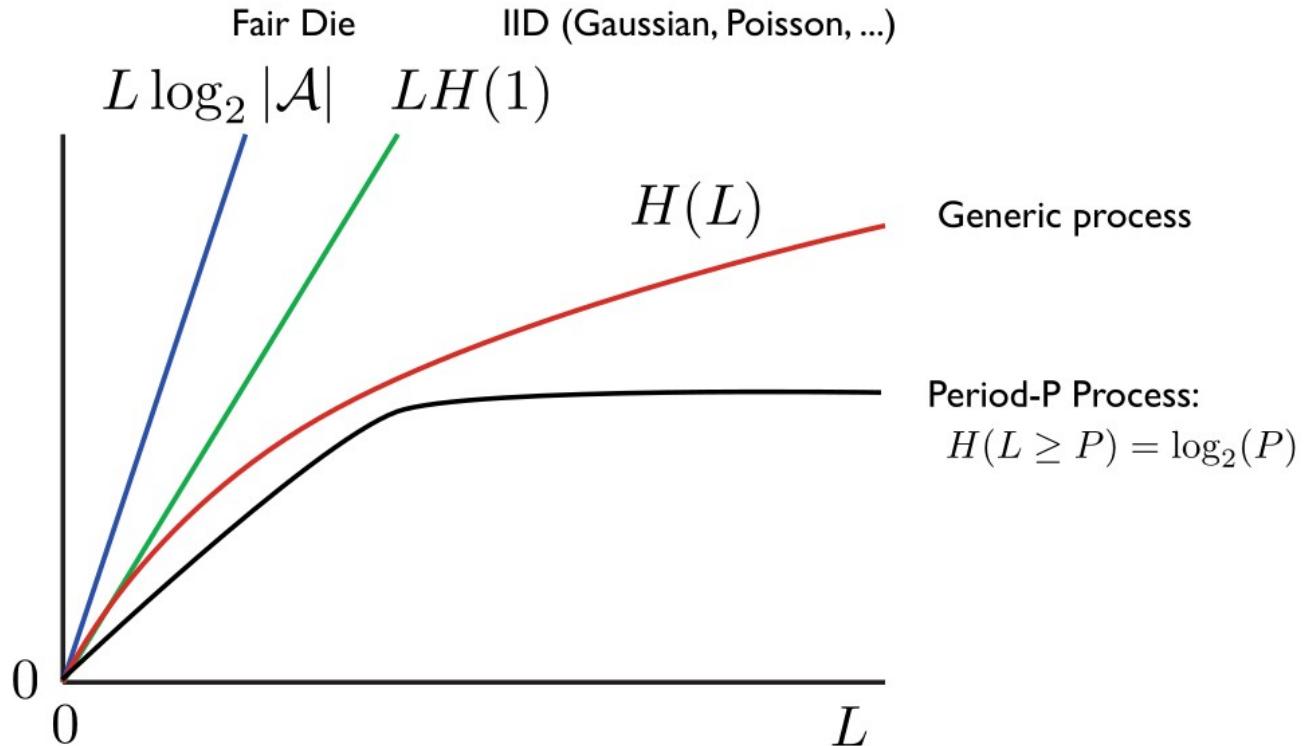
Asymptotic growth rate of entropy

Irreducible randomness of process

Average description length (per symbol) of process

Information in processes

Entropy Growth for Stationary Stochastic Processes ...
Block Entropy ...



Information in processes

Entropy Convergence:

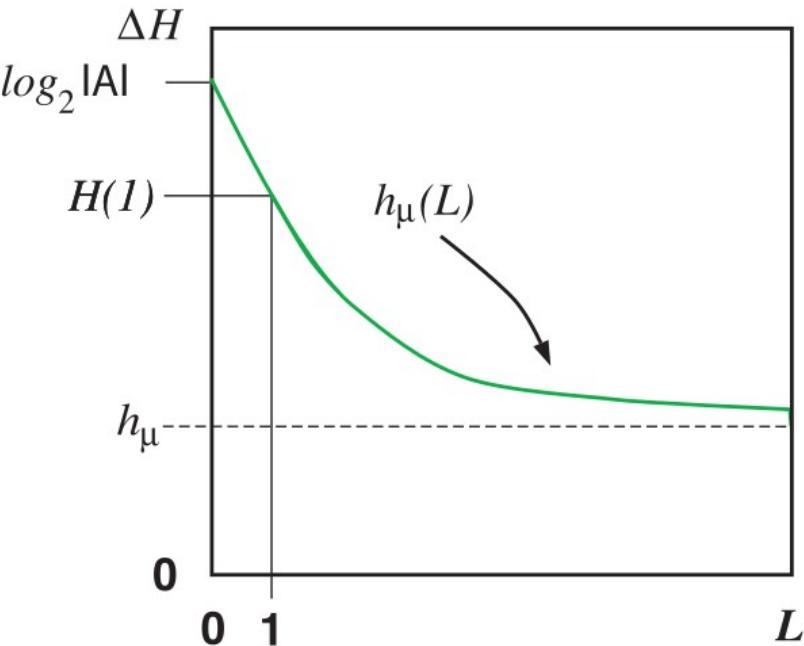
Length-L entropy rate estimate:

$$h_\mu(L) = H(L) - H(L-1)$$

$$h_\mu(L) = \Delta H(L)$$

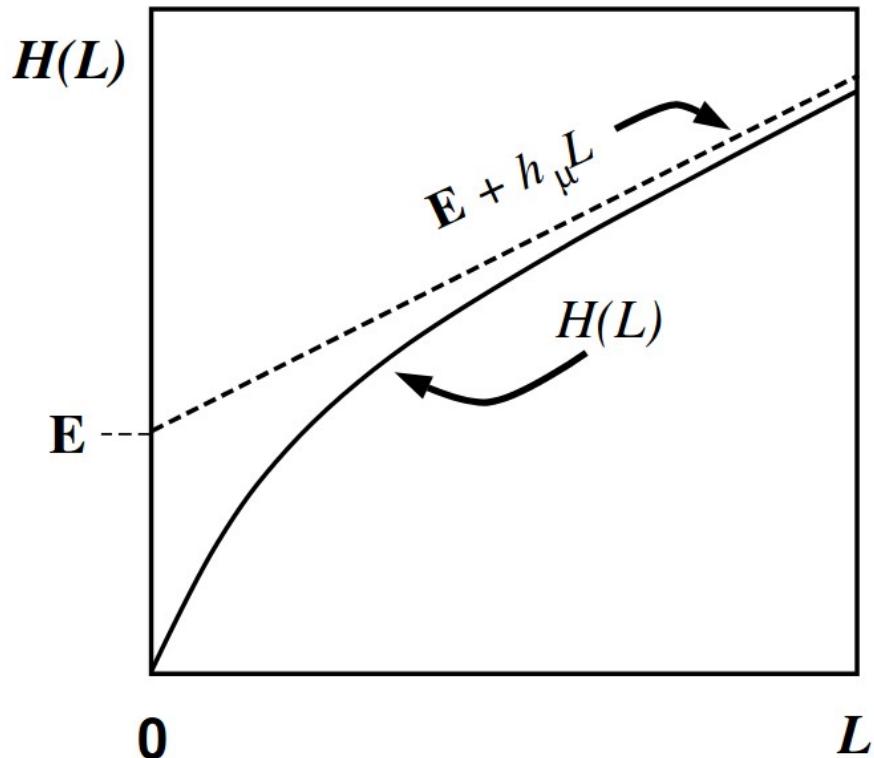
Monotonic decreasing:

$$h_\mu(L) \leq h_\mu(L-1)$$



Process appears less random
as account for longer correlations

Information in processes



E is the **excess entropy**: measures the redundancy arising from “correlations”

h_μ is the **entropy rate**: average entropy per symbol, given the past

Information in processes

Memory in Processes ...

Examples of Excess Entropy:

Fair Coin:

$$h_\mu = 1 \text{ bit per symbol}$$

$$\mathbf{E} = 0 \text{ bits}$$

Biased Coin:

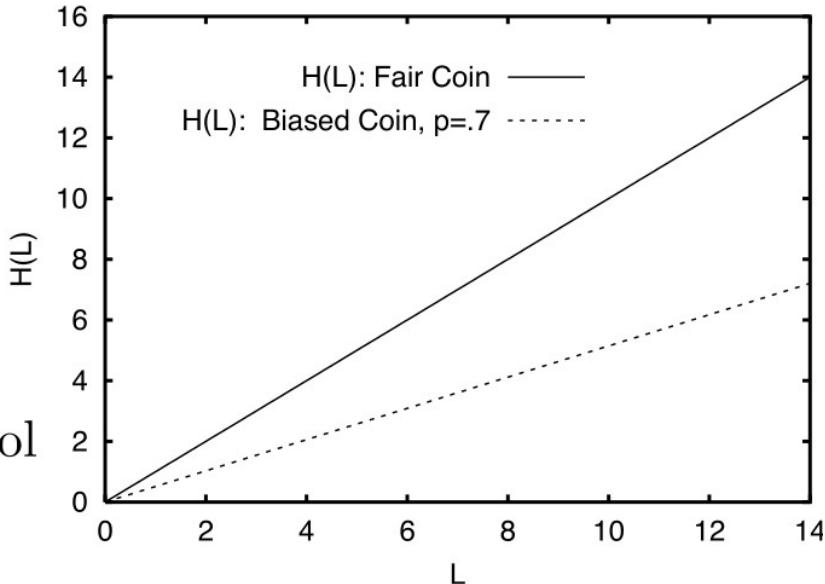
$$h_\mu = H(p) \text{ bits per symbol}$$

$$\mathbf{E} = 0 \text{ bits}$$

Any IID Process:

$$h_\mu = H(X) \text{ bits per symbol}$$

$$\mathbf{E} = 0 \text{ bits}$$

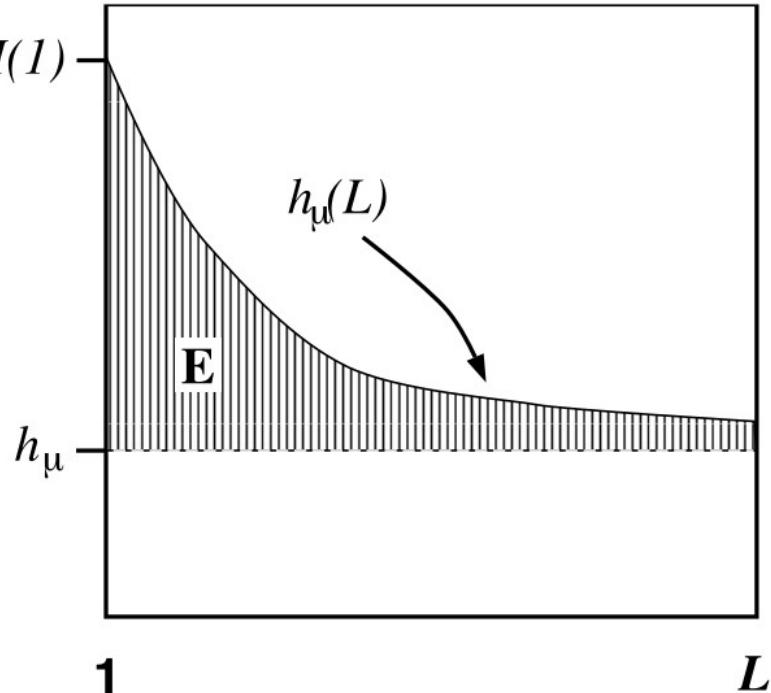


Information for processes

Excess Entropy:

As entropy convergence:

$$E = \sum_{L=1}^{\infty} [h_{\mu}(L) - h_{\mu}] \quad (\Delta L = 1 \text{ symbol})$$



Properties:

- (1) Units: $E = [\text{bits}]$
- (2) Positive: $E \geq 0$
- (3) Controls convergence to actual randomness.
- (4) Slow convergence \Leftrightarrow Correlations at longer words.
- (5) Complementary to entropy rate.

Information in processes

Motivation:

Previous: Measures of randomness of information source

Block entropy $H(L)$

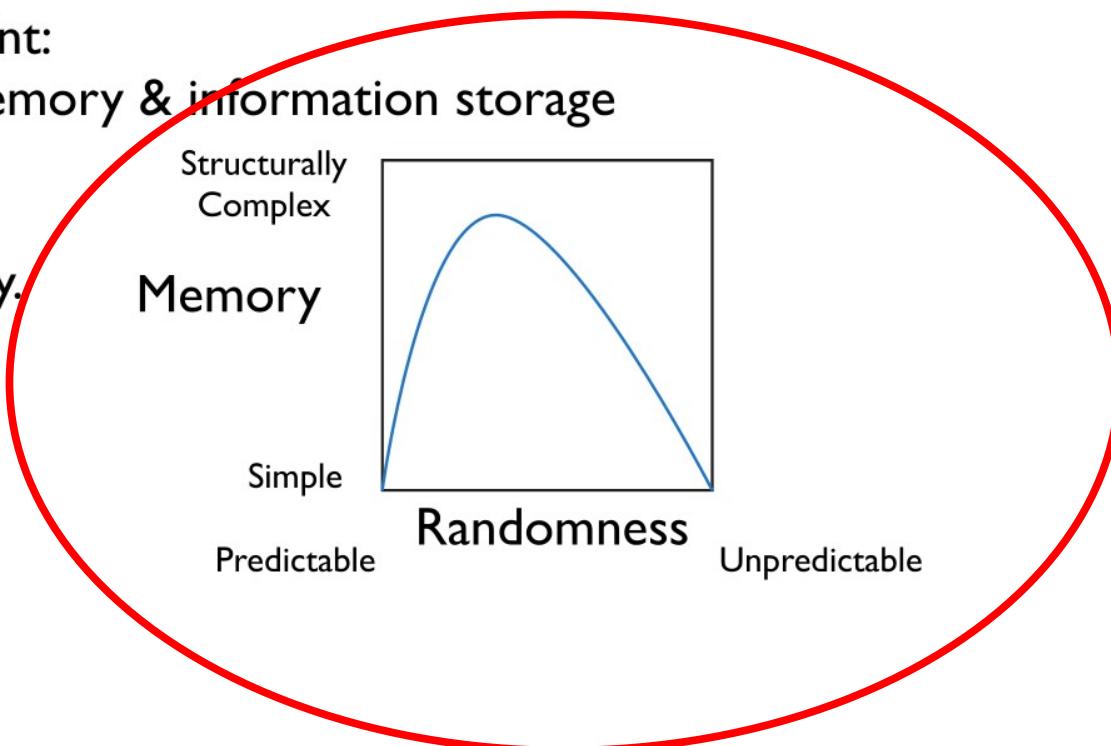
Entropy rate h_μ

Current target point:

Measures of memory & information storage

Big Picture:

Complementary.

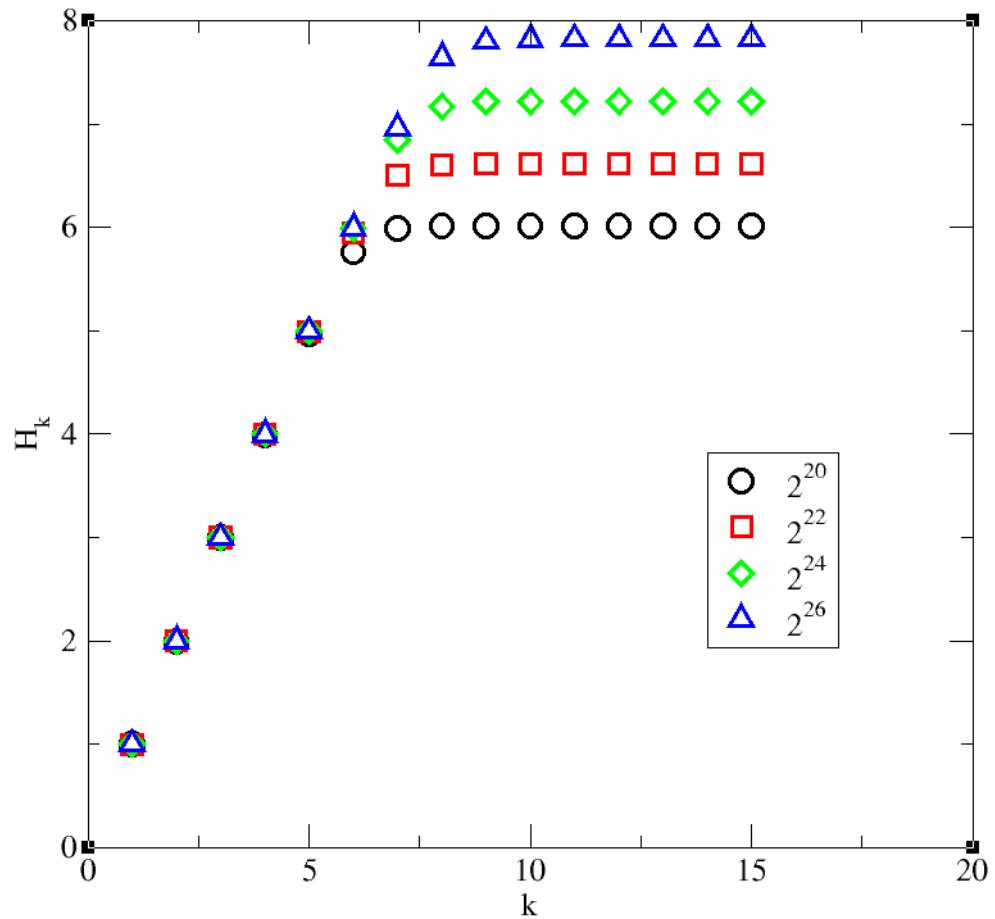


Examples

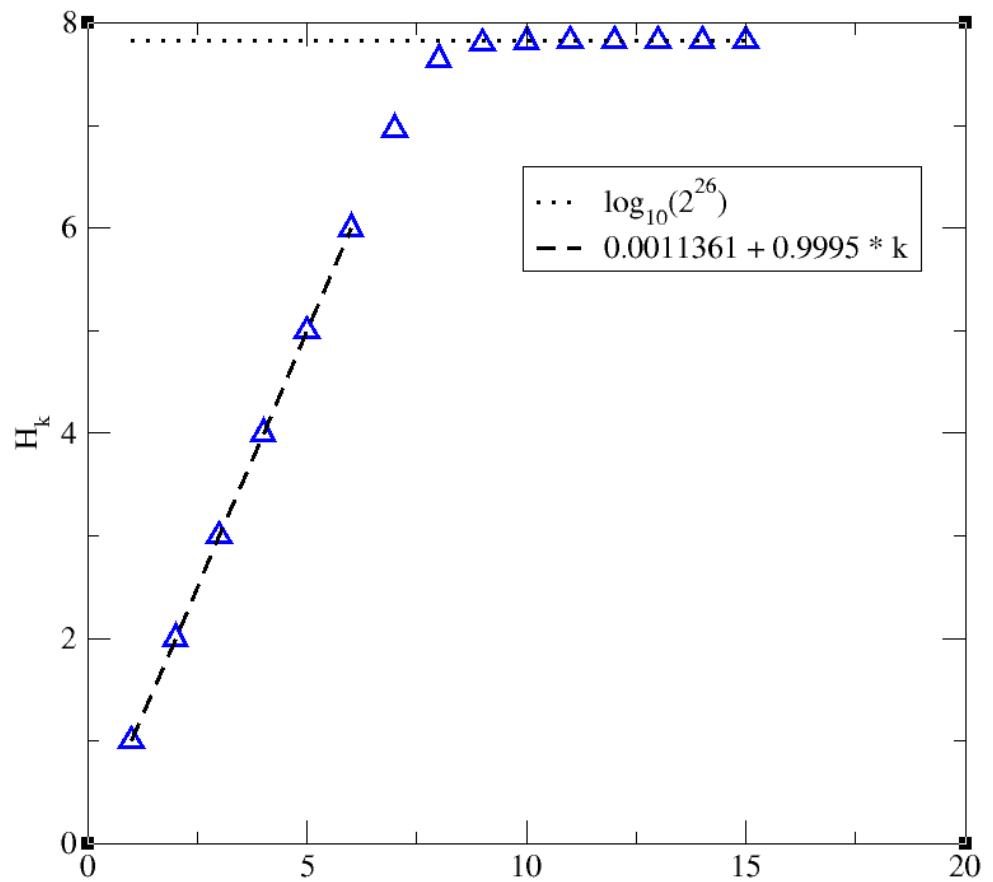
141592653589793238462643383279502884197169399375105820974944592307816406286208998628034825342117067982148086513282306647093844609550582
231725359408128481117450284102701938521105559644622948954930381964428810975665933446128475648233786783165271201909145648566923460348610
454326648213393607260249141273724587006606315588174881520920962829254091715364367892590360011330530548820466521384146951941511609433057
270365759591953092186117381932611793105118548074462379962749567351885752724891227938183011949129833673362440656643086021394946395224737
190702179860943702770539217176293176752384674818467669405132000568127145263560827785771342757789609173637178721468440901224953430146549
585371050792279689258923542019956112129021960864034418159813629774771309960518707211349999998372978049951059731732816096318595024459455
346908302642522308253344685035261931188171010003137838752886587533208381420617177669147303598253490428755468731159562863882353787593751
957781857780532171226806613001927876611195909216420198938095257201065485863278865936153381827968230301952035301852968995773622599413891
249721775283479131515574857242454150695950829533116861727855889075098381754637464939319255060400927701671139009848824012858361603563707
660104710181942955596198946767837449448255379774726847104047534646208046684259069491293313677028989152104752162056966024058038150193511
253382430035587640247496473263914199272604269922796782354781636009341721641219924586315030286182974555706749838505494588586926995690927
210797509302955321165344987202755960236480665499119881834797753566369807426542527862551818417574672890977772793800081647060016145249192
17321721477235014144197356854816136115735255213347574184946843852332390739414333454776241686251898356948556209921922184272550254256887
671790494601653466804988627232791786085784383827967976681454100953883786360950680064225125205117392984896084128488626945604241965285022
210661186306744278622039194945047123713786960956364371917287467764657573962413890865832645995813390478027590099465764078951269468398352
595709825822620522489407726719478268482601476990902640136394437455305068203496252451749399651431429809190659250937221696461515709858387
41059788595772975498930161753928468138268683868942774155991855925245953959431049972524680845987273644695848653836736222626099124608051
24388439045124413654976278079771569143599770012961608944169486855584840635342207222582848864815845602850601684273945226746768895252138
522549954666727823986456596116354886230577456498035593634568174324112515076069479451096596094025228879710893145669136867228748940560101
50330861792868092087476091782493858900971490967598526136554978189312978482168299894872265880485756401427047755132379641451523746234364
542858444795265867821051141354735739523113427166102135969536231442952484937187110145765403590279934403742007310578539062198387447808478
489683321445713868751943506430218453191048481005370614680674919278191197939952061419663428754440643745123718192179998391015919561814675
142691239748940907186494231961567945208095146550225231603881930142093762137855956638937787083039069792077346722182562599661501421503068
038447734549202605414665925201497442850732518666002132434088190710486331734649651453905796268561005508106658796998163574736384052571459
102897064140110971206280439039759515677157700420337869936007230558763176359421873125147120532928191826186125867321579198414848829164470
609575270695722091756711672291098169091528017350671274858322287183520935396572512108357915136988209144421006751033467110314126711136990
865851639831501970165151168517143765761835155650884909989859982387345528331635507647918535893226185489632132933089857064204675259070915
481416549859461637180270981994309924488957571282890592323326097299712084433573265489382391193259746366730583604142813883032038249037589
852437441702913276561809377344403070746921120191302033038019762110110044929321516084244485963766983895228684783123552658213144957685726
243344189303968642624341077322697802807318915441101044682325271620105265227211166039666557309254711055785376346682065310989652691862056
476931257058635662018558100729360659876486117910453348850346113657686753249441668039626579787718556084552965412665408530614344431858676

Examples

Pi digits

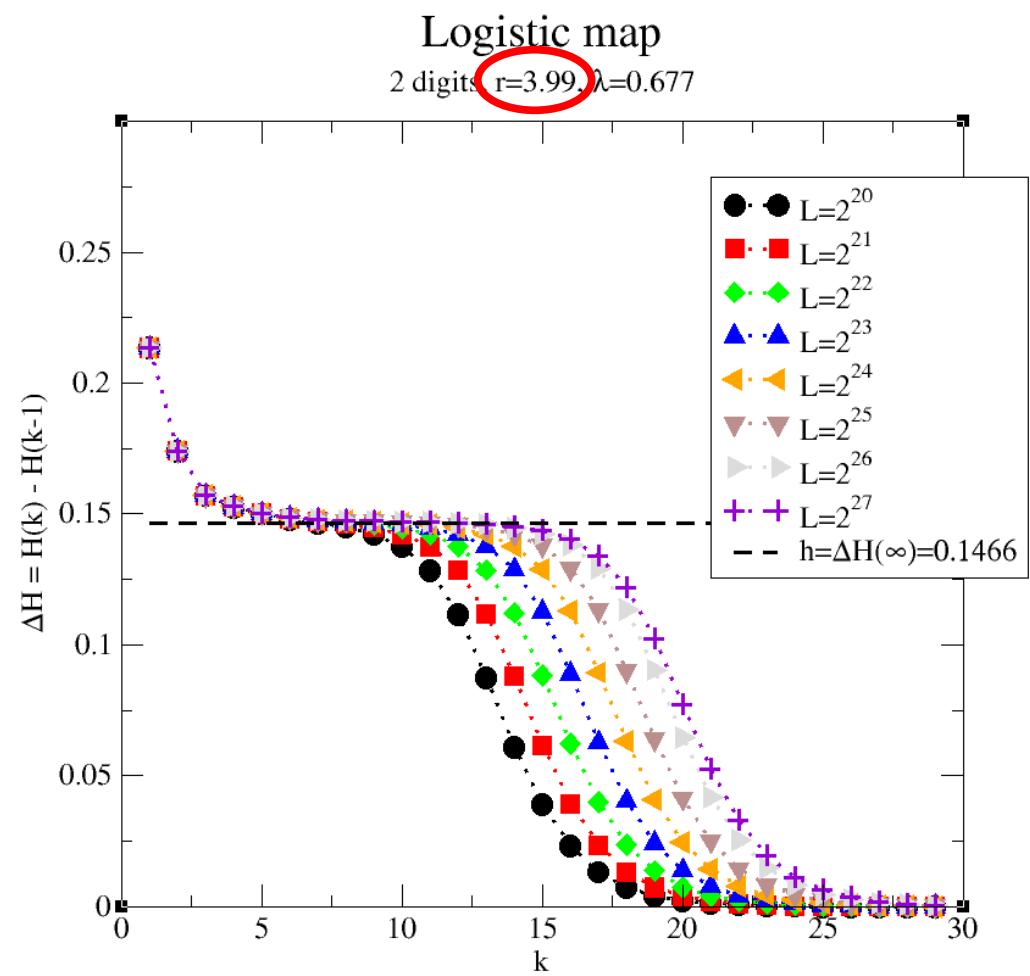
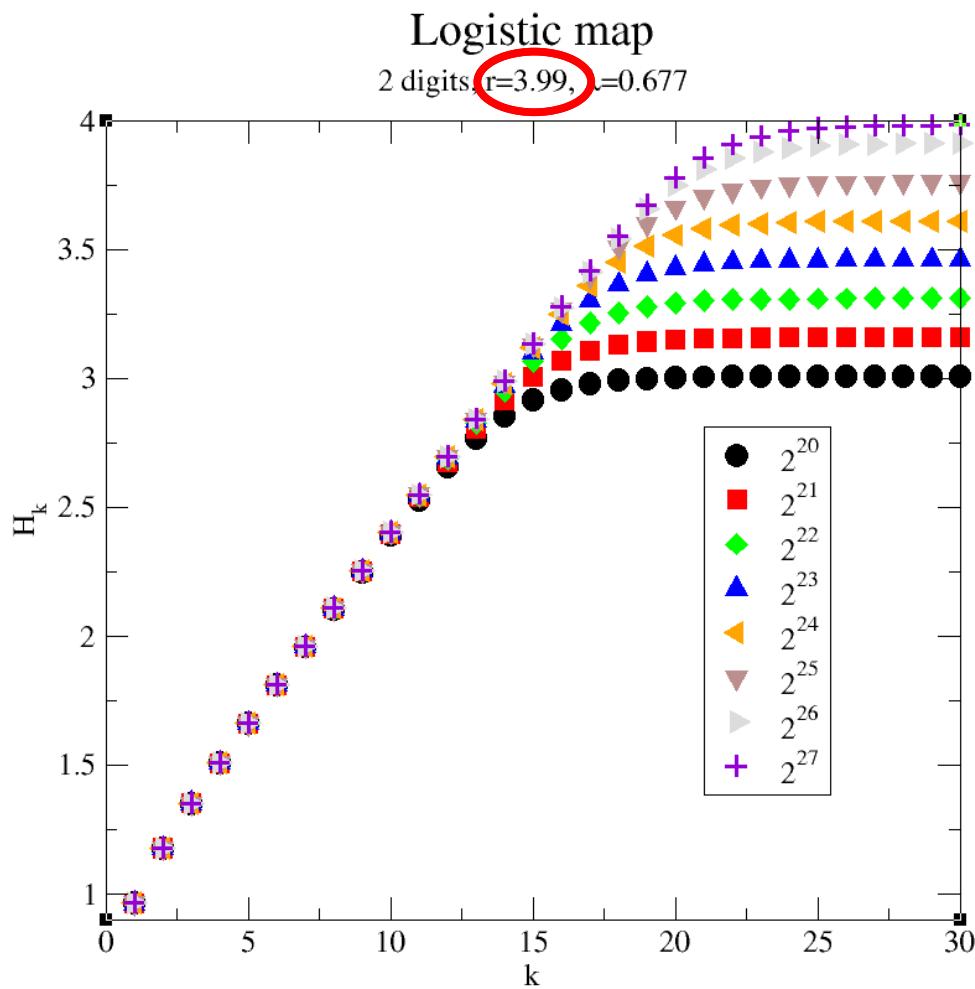


Pi digits

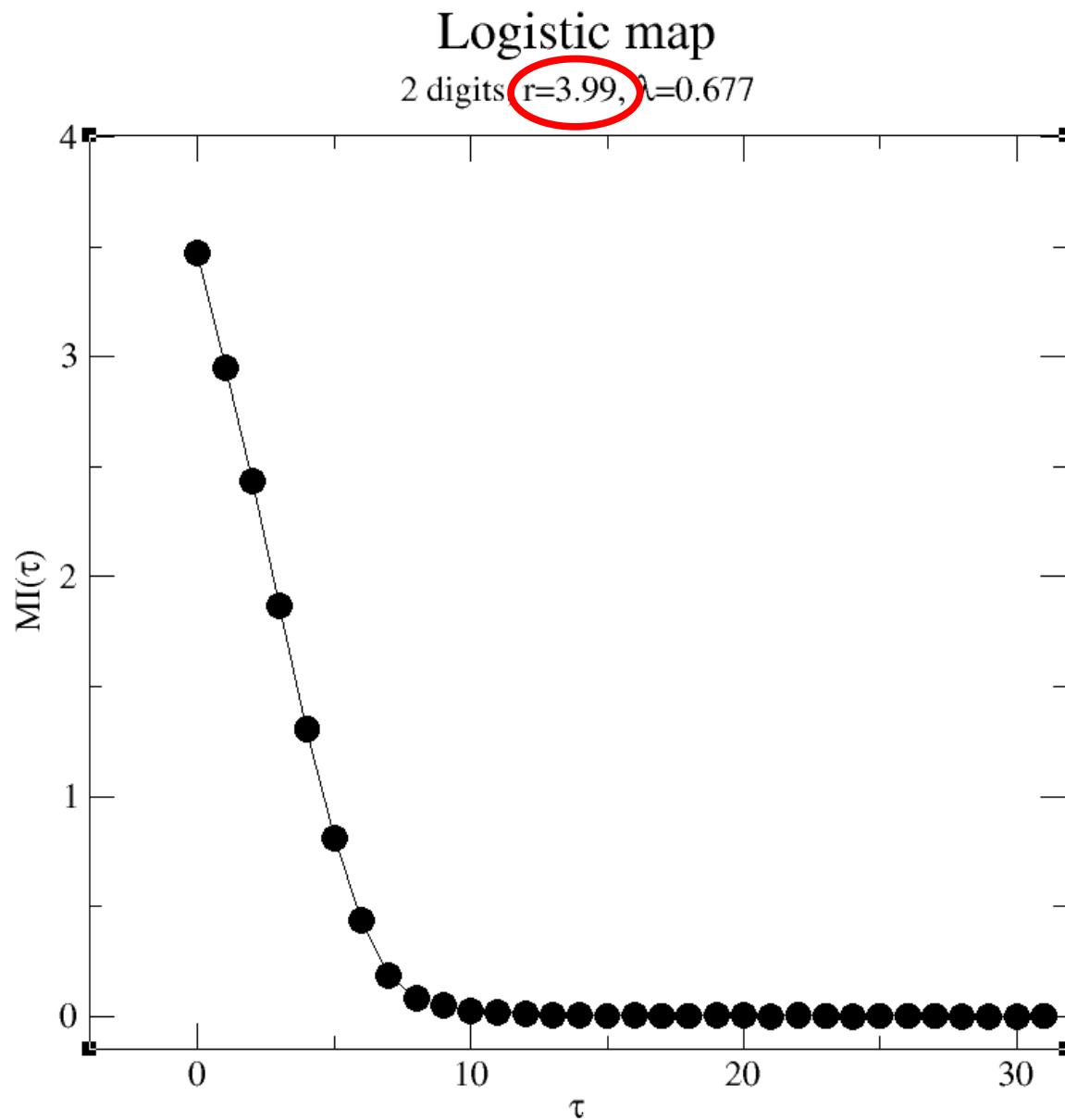


$$\log_2 x = \log_{10} x / \log_{10} 2 \approx 3.322 \log_{10} x$$

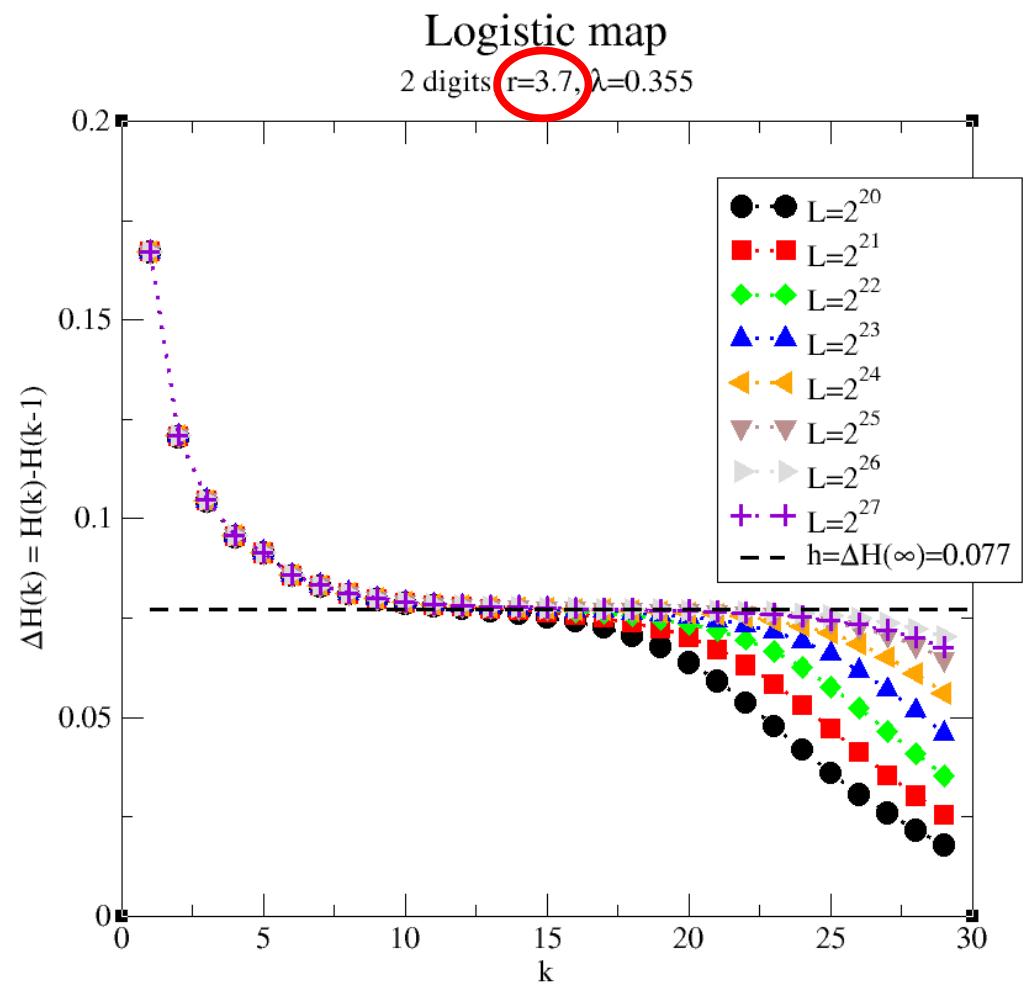
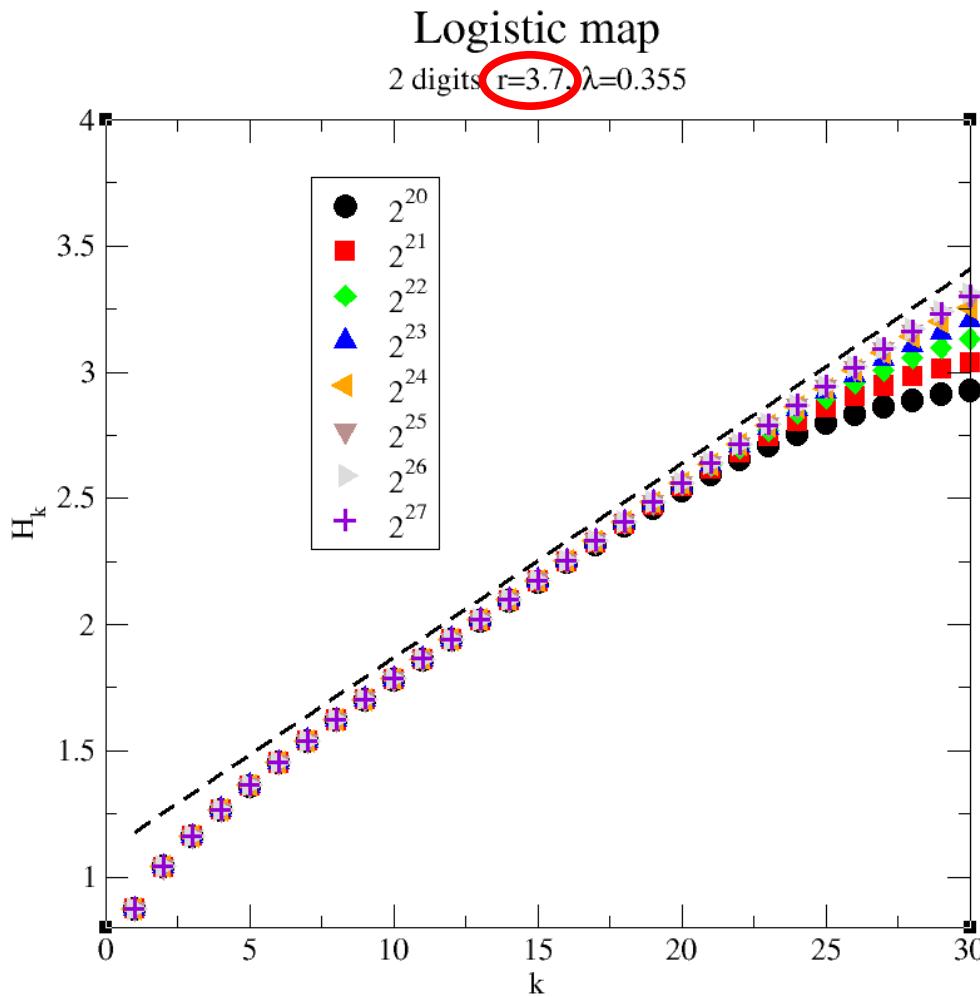
Examples



Examples

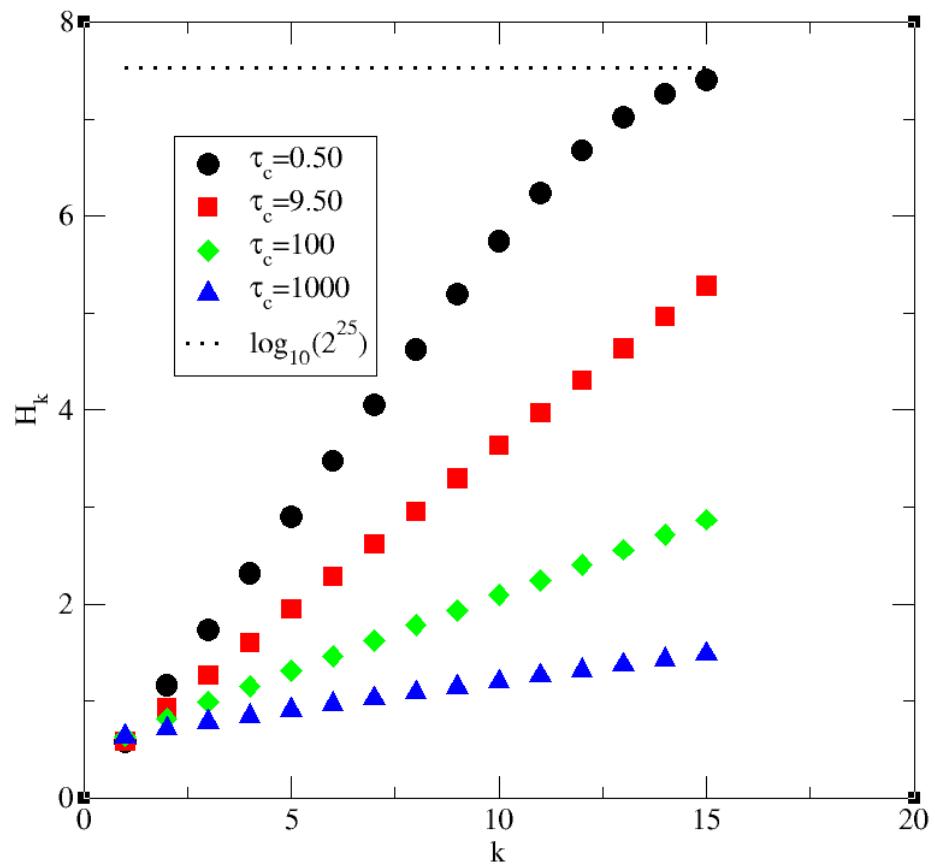


Examples

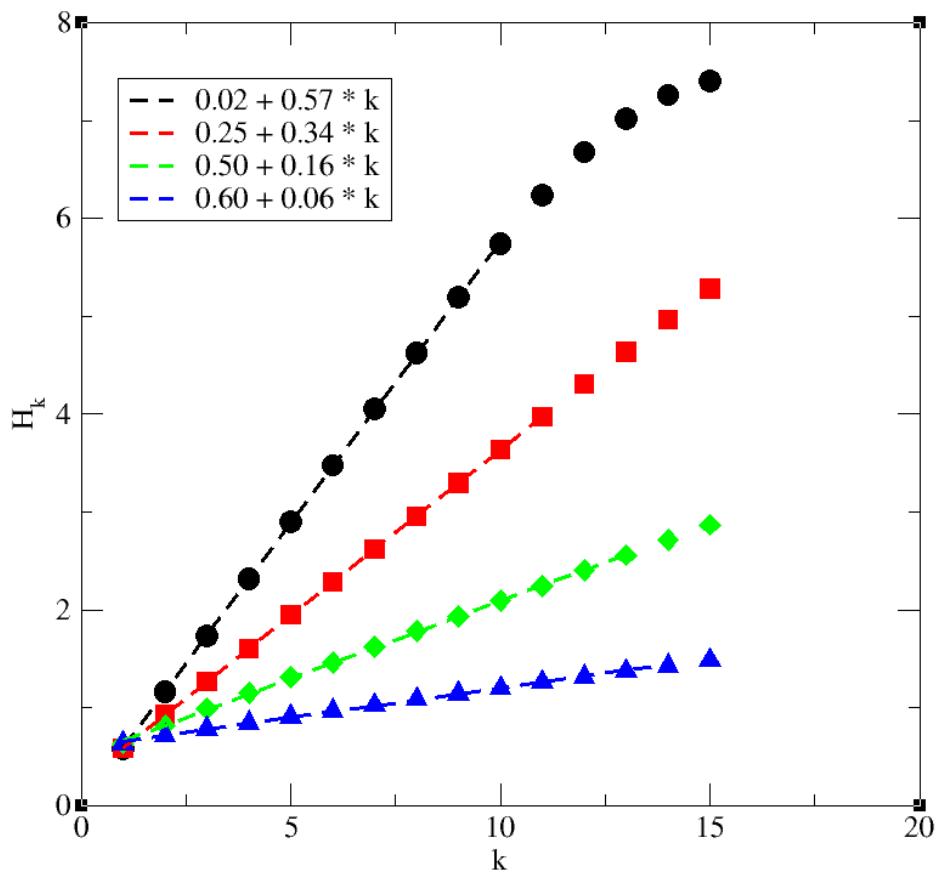


Examples

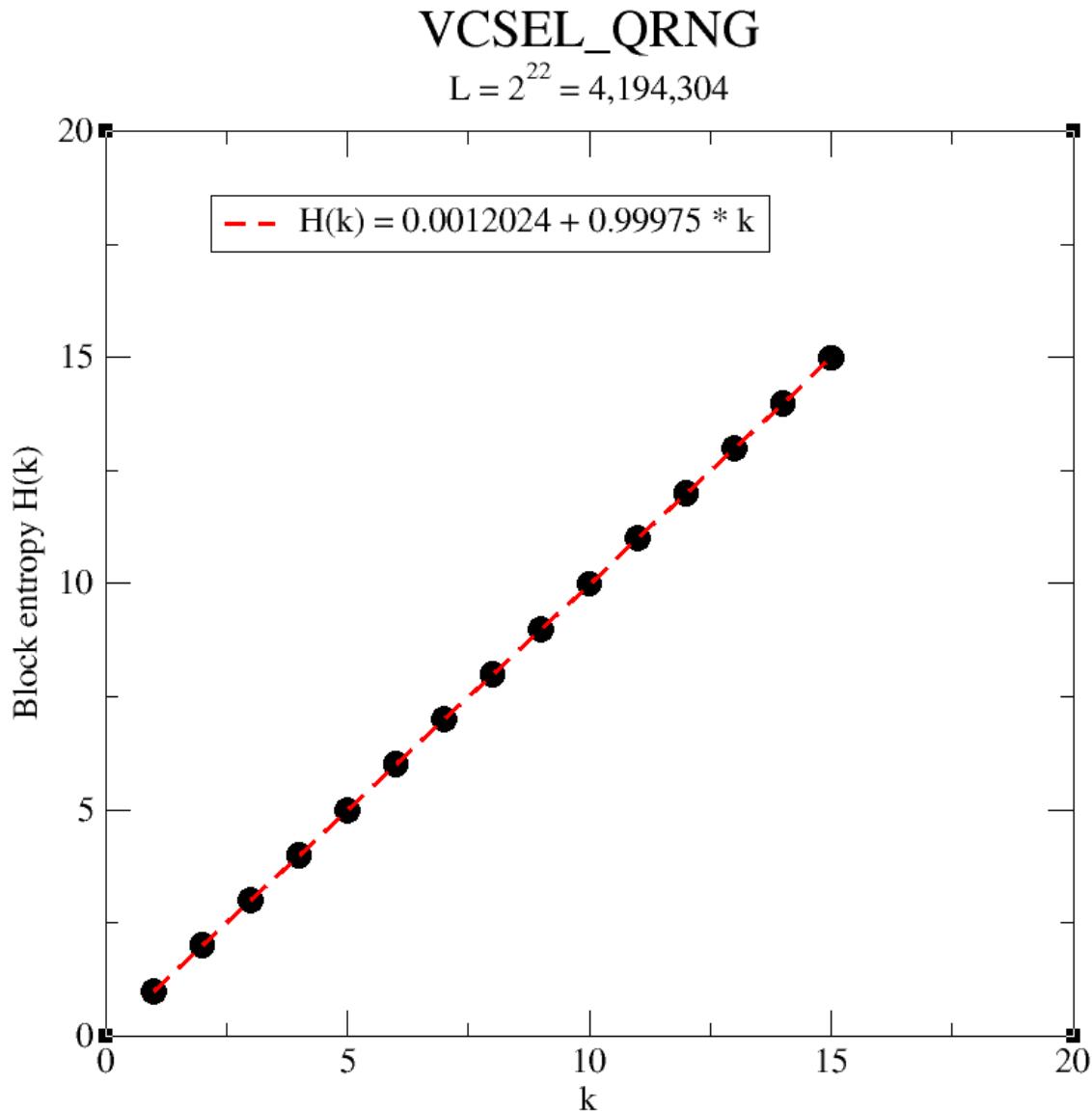
Exponential decay of correlations



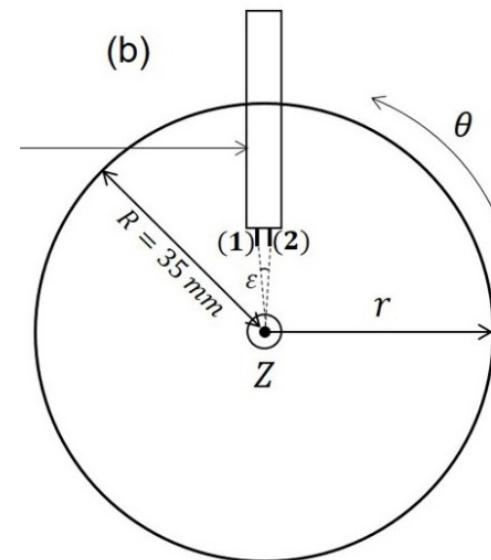
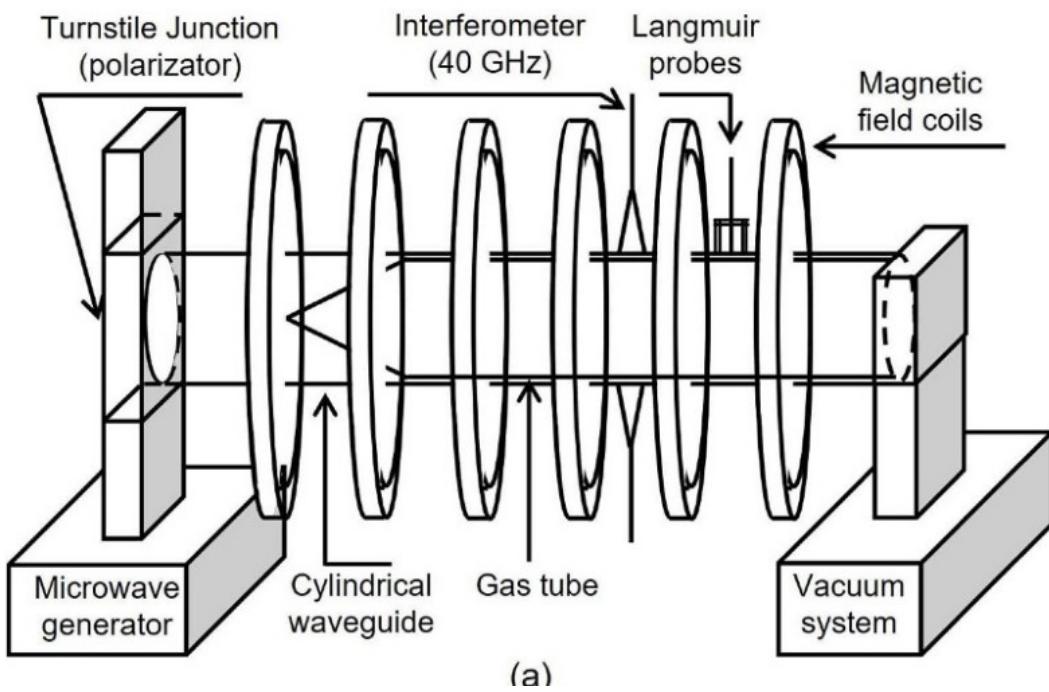
Exponential decay of correlations



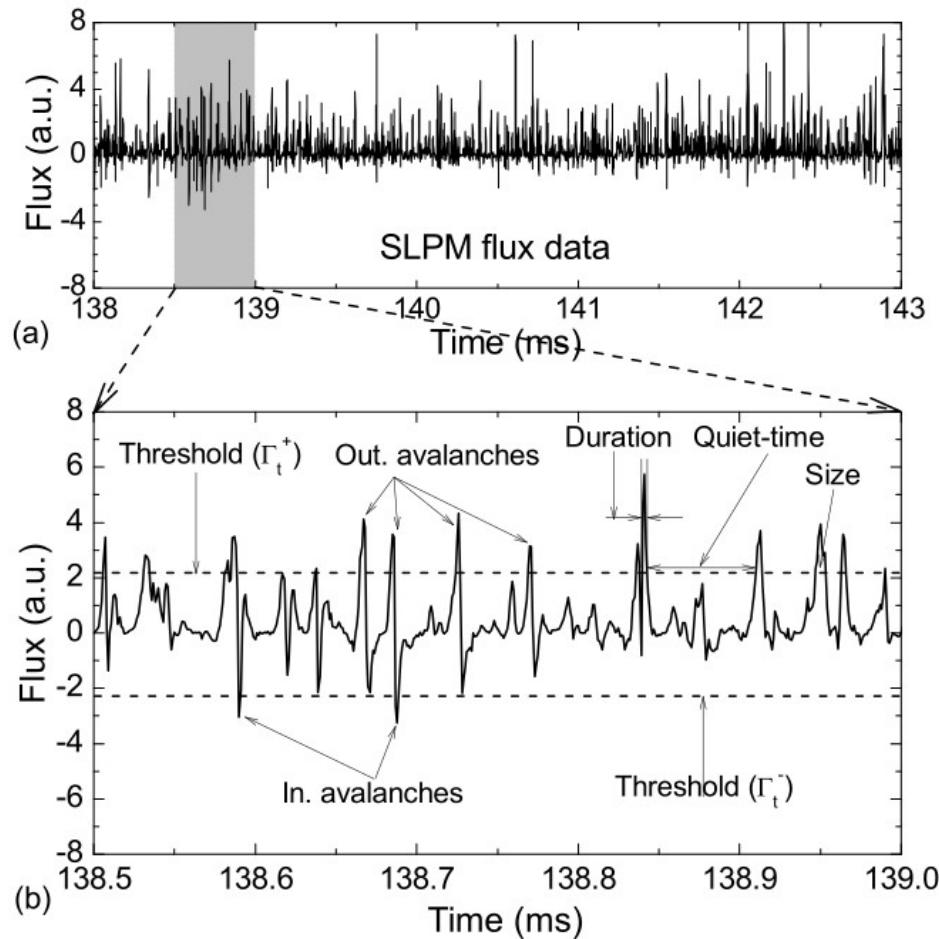
Examples



Examples

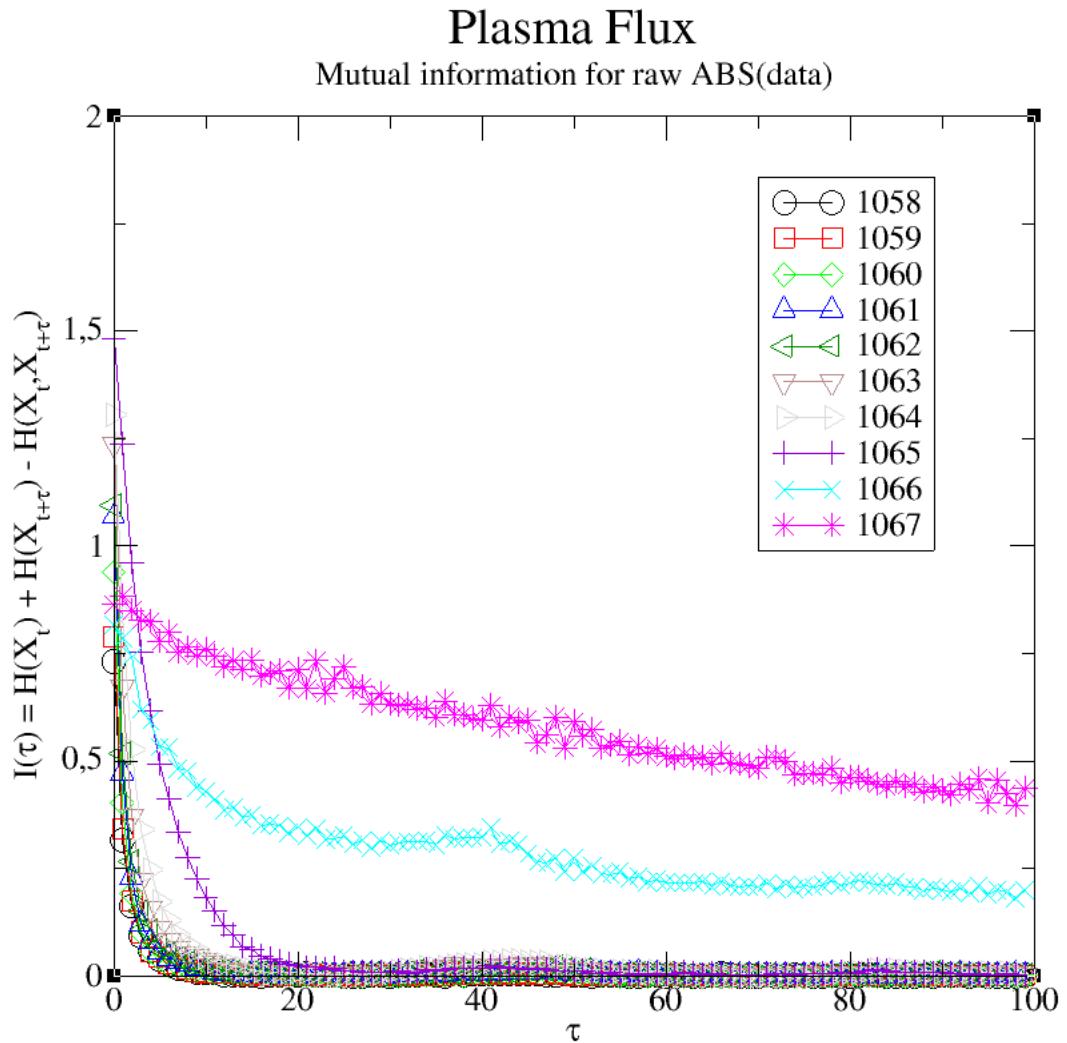


Examples

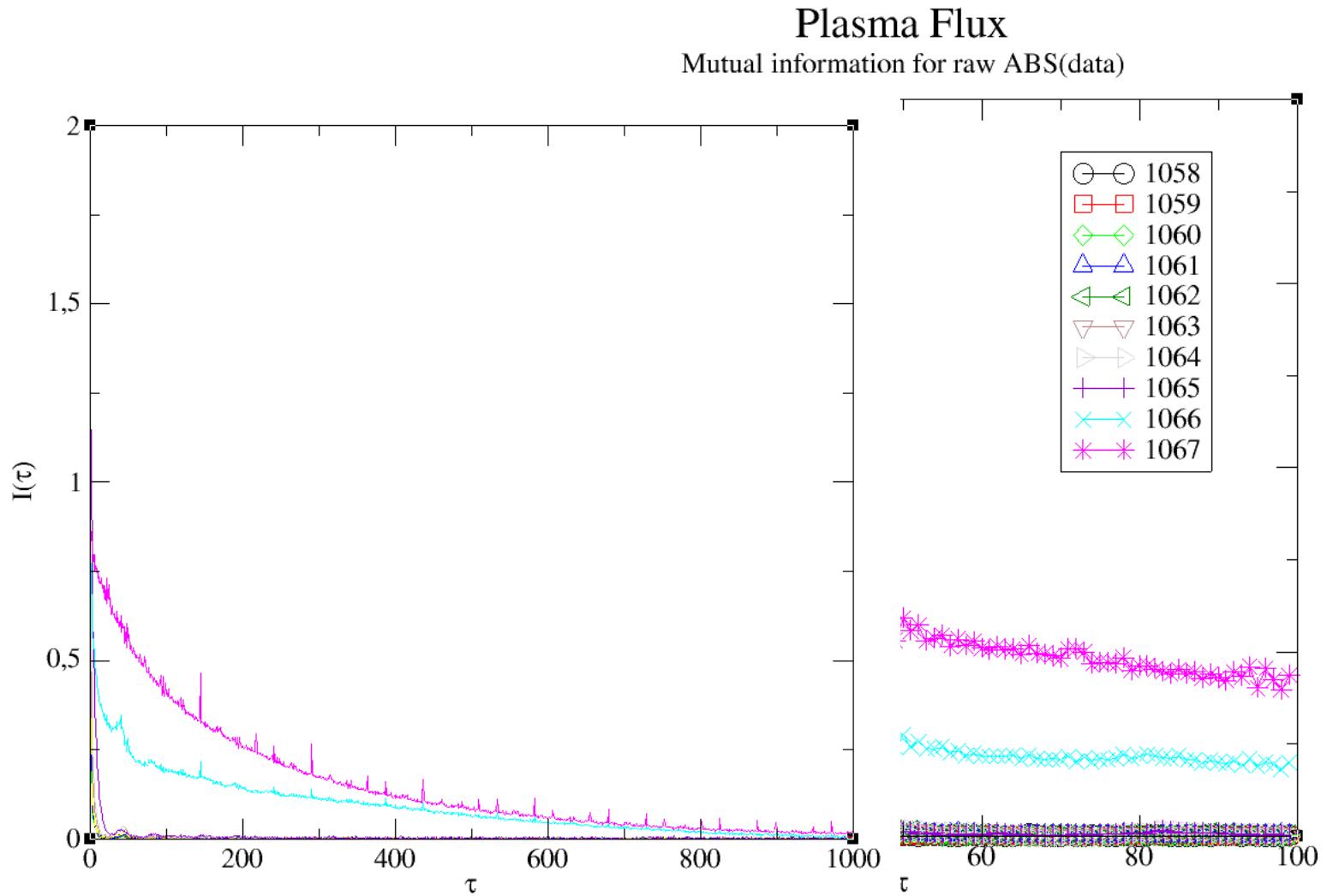


200.000 data points

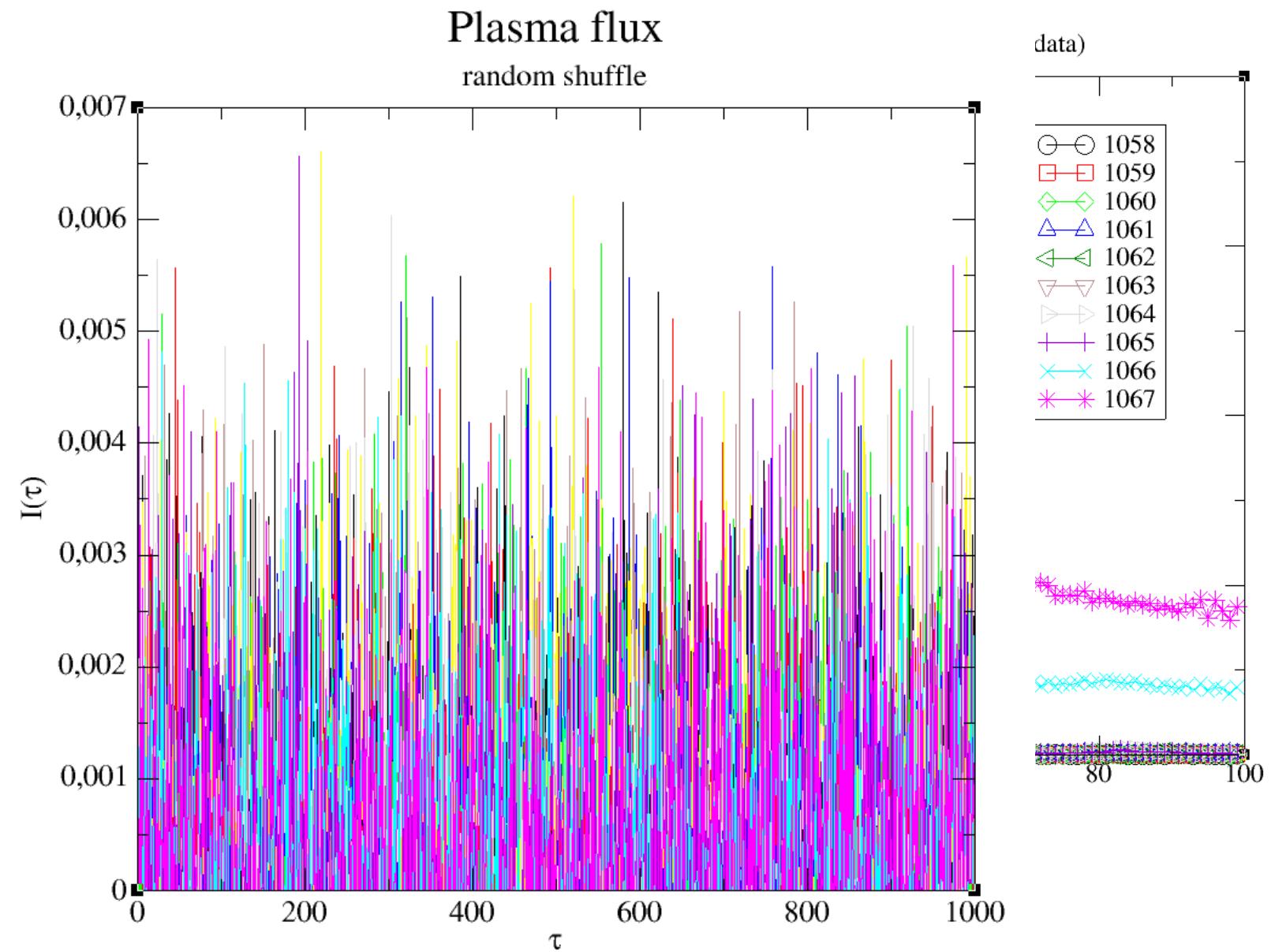
Examples



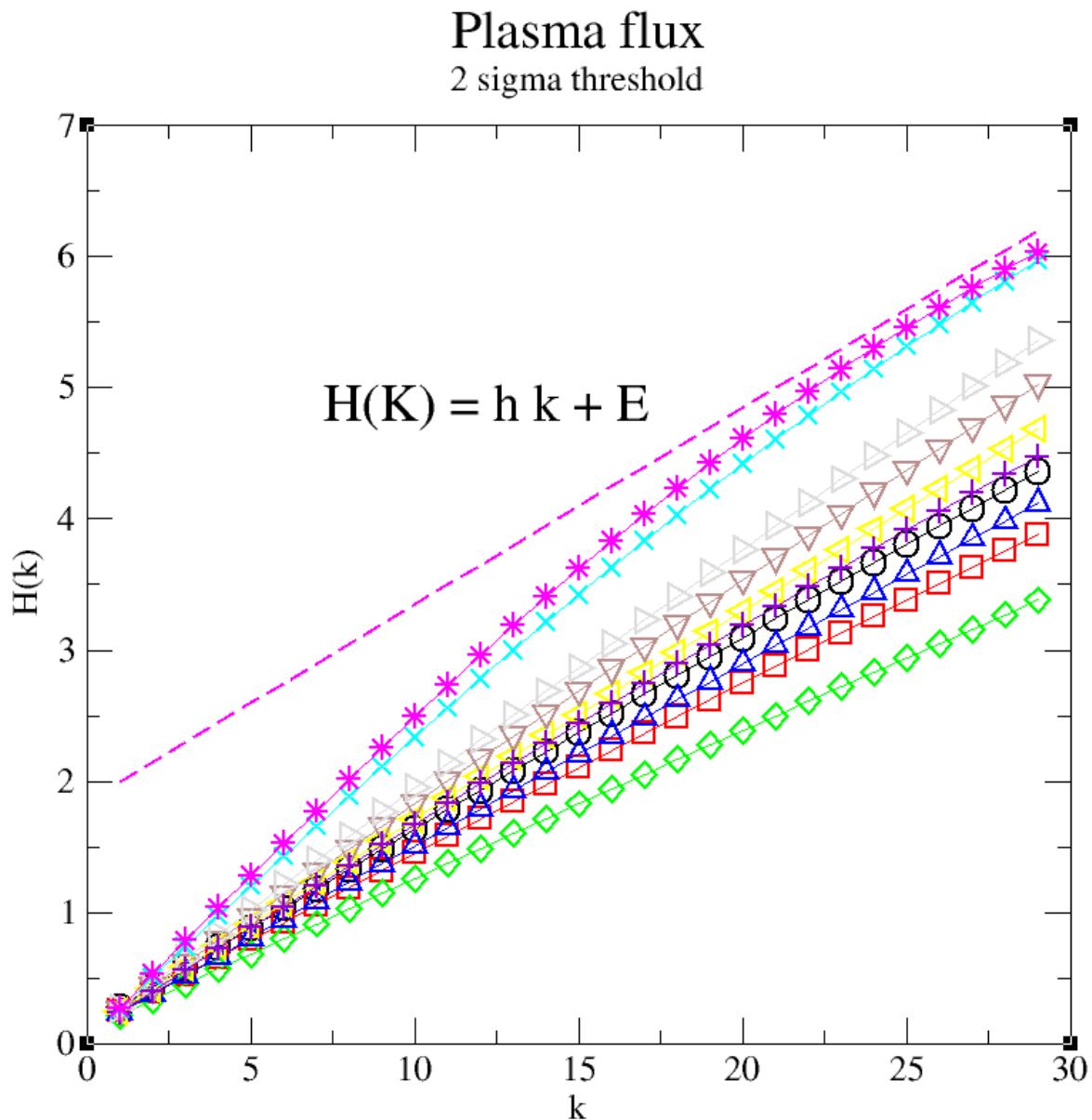
Examples



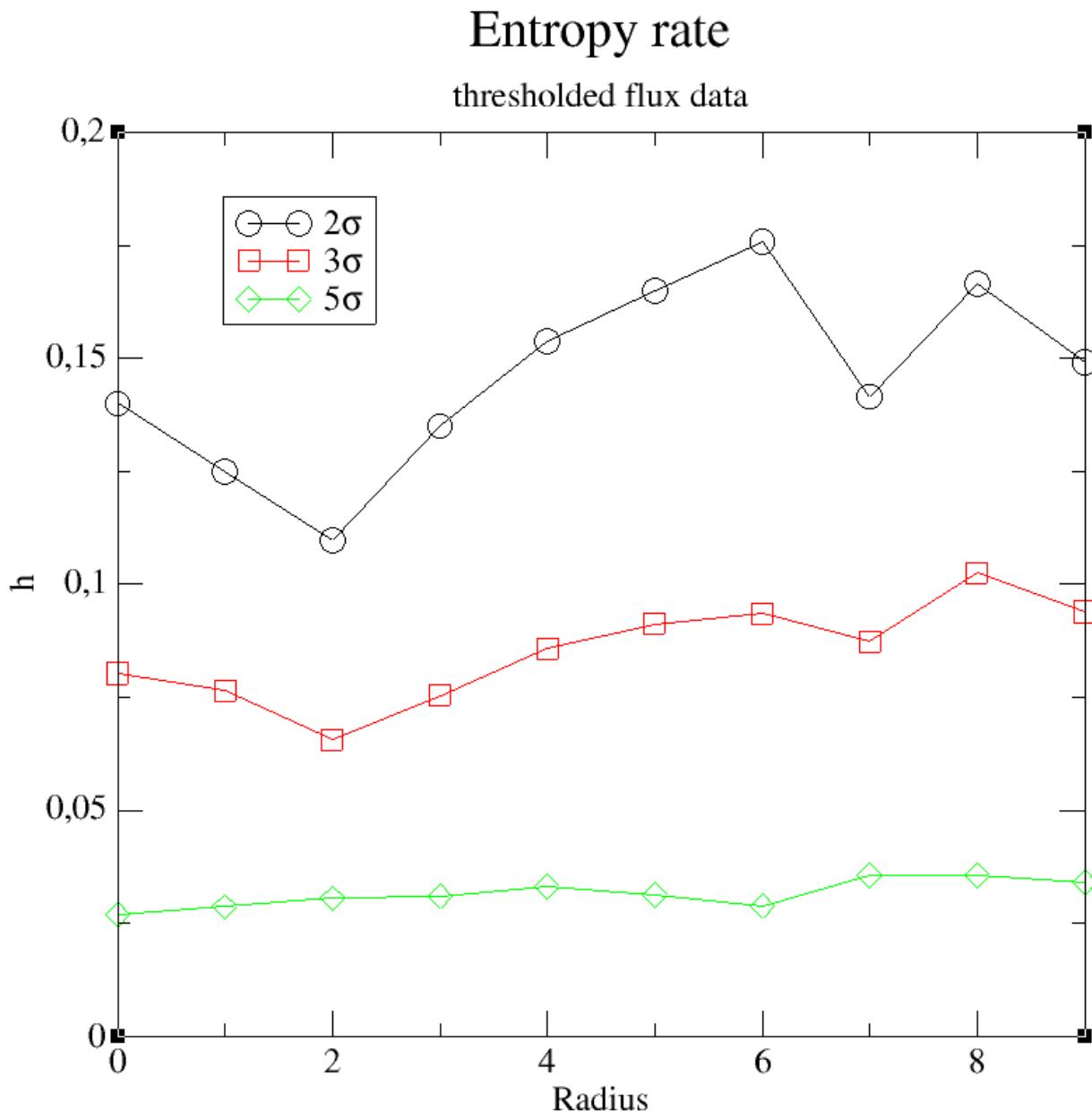
Examples



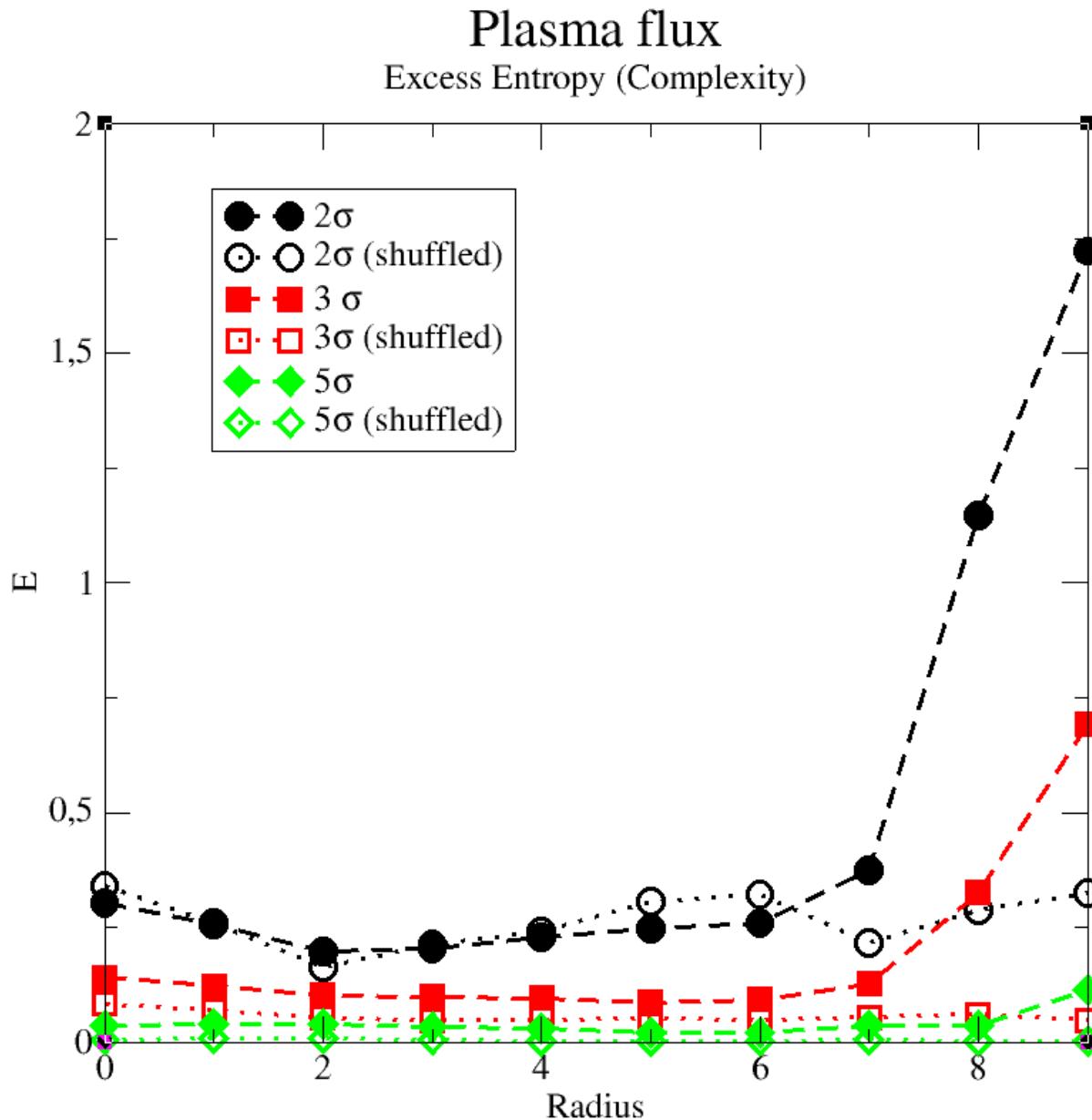
Examples



Examples



Examples



Maximum entropy principle

Maximum entropy principle

- Maximum entropy principle arose in statistical mechanics
- If nothing is known about a distribution except that it belongs to a certain class
- Distribution with the largest entropy should be chosen as the default
- Motivation:
 - Maximizing entropy minimizes the amount of prior information built into the distribution
 - Many physical systems tend to move towards maximal entropy configurations over time

Maximum entropy principle

Physics

- Temperature of a gas corresponds to the average kinetic energy of the molecules in the gas

$$\sum_i p_i \frac{1}{2} v_i^2 m_i$$

- Distribution of velocities in the gas at a given temperature
- this distribution is the maximum entropy distribution under the temperature constraint: Maxwell-Boltzmann distribution
- corresponds to the macrostate that has the most micro states

Maximum entropy principle

Formulation

- Maximize entropy

$$H(p) = - \sum_{i=1}^n p_i \log p_i$$

- Subject to

$$p_i \geq 0 \tag{1}$$

$$\sum_{i=1}^n p_i = 1 \tag{2}$$

$$\sum_{i=1}^n p_i r_{ij} = \alpha_j, \text{ for } 1 \leq j \leq m \tag{3}$$

Maximum entropy principle

Maximum entropy distribution

- Form Lagrangian

$$J(p) = - \sum_{i=1}^n p_i \log p_i + \lambda_0 \left(\sum_{i=1}^n p_i - 1 \right) + \sum_{j=1}^m \lambda_j \left(\sum_{i=1}^n p_i r_{ij} - \alpha_j \right)$$

- Take derivative with respect to p_i : $-1 - \log p_i + \lambda_0 + \sum_{j=1}^m \lambda_j r_{ij}$
- Set this to 0, and solution is *maximum entropy distribution*

$$p_i^* = \frac{e^{\sum_{j=1}^m \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- $\lambda_0, \lambda_1, \dots, \lambda_m$ are chosen such that $\sum_i p_i^* = 1$, and $\sum_i p_i^* r_{ij} = \alpha_j$.

Maximum entropy principle

Maximum entropy classifier

- In some fields of machine learning, multinomial logic model is refer to as a maximum entropy classifier
- Minimizes the amount of prior information built into the distribution
- X_i : feature vector, β_k : vector of weights for outcome k , Y_k : random outcome, $k = 1, \dots, K$
- Takes the form

$$p(Y_i = k) = \frac{e^{\beta_k^\top X_i}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell^\top X_i}}, \quad k = 1, \dots, K-1.$$

Summary

- Maximizing entropy minimizes the amount of prior information built into unknown distribution
- Maximum entropy distribution can be found explicitly

$$p_i^* = \frac{e^{\sum_{j=1}^m \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- Maximum entropy principle widely used