EXAMPLE OF H

FANO ENCODING

ENTROPY AND NUMBER OF BITS

LABEL INVARIANCE

$H(f(x))$

ALTERNATIVE ENTROPIES FOR ANOMALY DETECTION

PERMUTATION ENTROPY

SAMPLE ENTROPY

EXERCISE 3

JOINT ENTROPY

$H(X,Y) = H(x) + H(Y)$ FOR S.I. X Y

$H(X,Y) \leq H(x) + H(Y)$

CONDITIONAL ENTROPY

CHAIN RULE

MEANING OF $H(x|y)$

$0 \leq H(x|y) \leq H(x)$

INFORMATION GAIN

INTERPRETATION - FIGURE

FORMULAS

KULLBACK - LEIBLER DIVERGENCE

EXAMPLE

x_0 x_1

	p_0	p_1	h_0	h_1	H
→	0.5	0.5	1	1	1
→	0.1	0.9	3.32	0.15	0.46
→	0.01	0.99	6.64	0.015	0.08

FANO ENCODING

R.V. X Ω_X $P(x)$

WE WANT TO ASSIGN TO EACH OUTCOME

$n \in \Omega_X$ A BINARY VECTOR

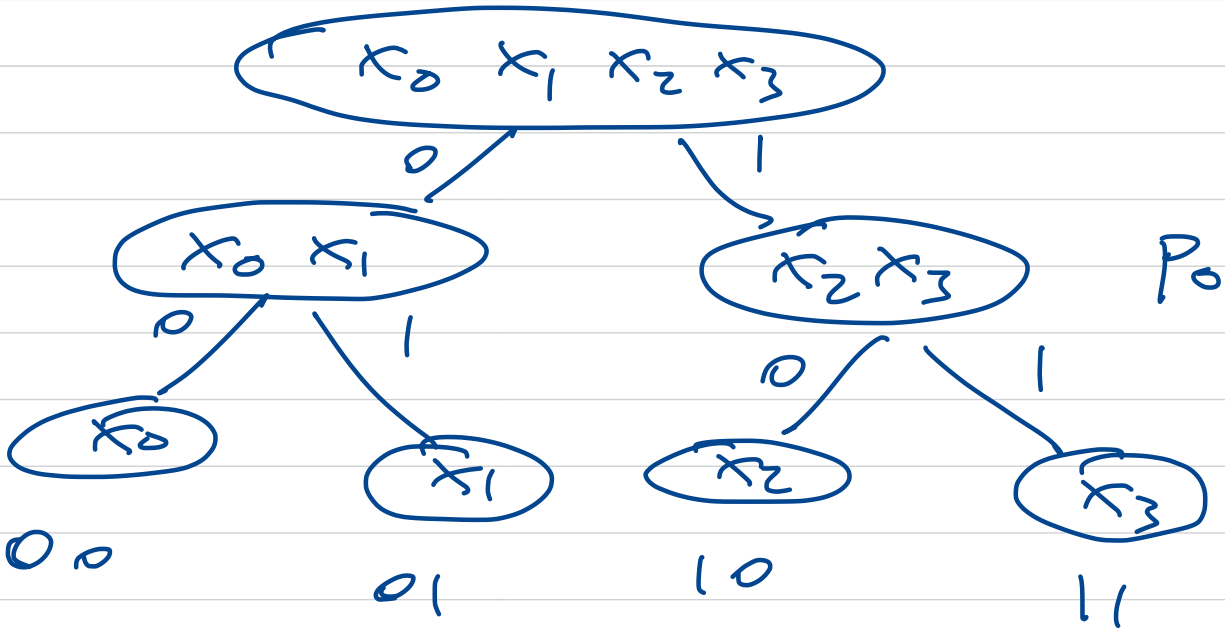
$$\Omega_X = \{ x_0 \quad x_1 \quad x_2 \quad x_3 \}$$

$$00 \quad 01 \quad 10 \quad 11$$

IDEA: INSTEAD OF USING FIXED LENGTH CODES WE COULD USE CODES WITH DIFF. LENGTH TRYING TO ASSIGN SMALL CODES (FEW BITS) TO MOST PROBABLE SYMBOLS

FANO ENCODER

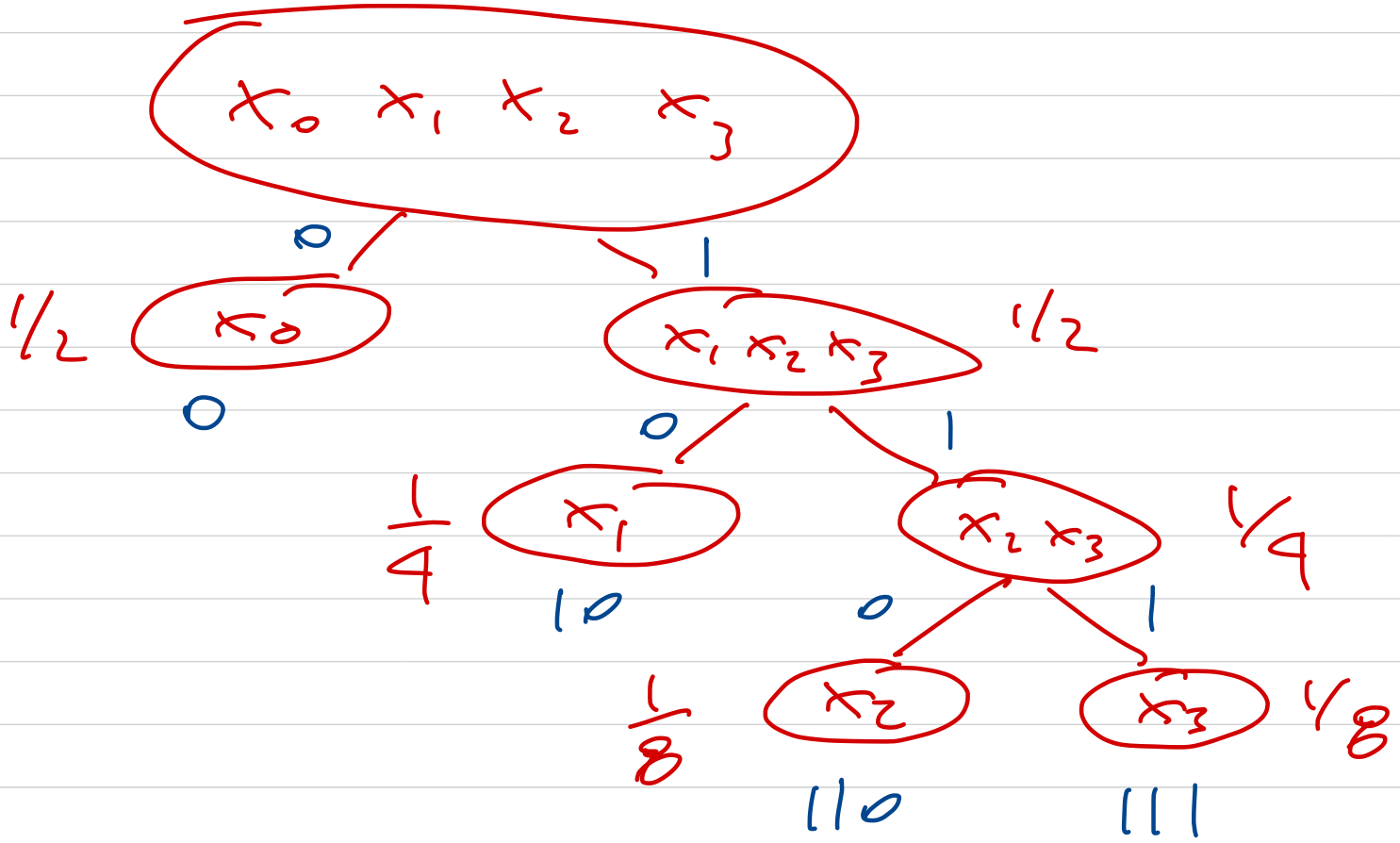
(NOT OPTIMAL, HUFFMAN ENCODER IS OPTIMAL → STUDIED IN SECTION 2)



WE DIVIDE EACH SET INTO TWO SETS SUCH THAT THE SUM OF THEIR PROB. IS AS CLOSE AS POSSIBLE

x_0	x_1	x_2	x_3
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

x_0	x_1	x_2	x_3
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
0	10	110	111



x_1	x_2	x_3	x_4
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
0	10	110	111
1	2	3	3
n_1	n_2	n_3	n_4

$$H(x) = \sum_{i=1}^{4} p_i \log_2 \frac{1}{p_i} =$$

$$= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 = 1.75$$

$$\bar{n} = p_1 n_1 + p_2 n_2 + p_3 n_3 + p_4 n_4$$

AVERAGE # OF BITS

$$= \frac{1}{2} + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3$$

$$= 1.75 \leq 2$$

OBTAINED WITH FIXED LENGTH CODE

IN GENERAL THIS PROPERTY HOLDS (SECTION 2)

$$H(x) \leq \tilde{m} \leq H(x) + 1$$

FANO ALGORITHM PRODUCES CODES SATISFYING THE PREFIX CONDITION

x_0 x_1 x_2 x_3 ← NO. PREFIX CODE

10 1 101 100

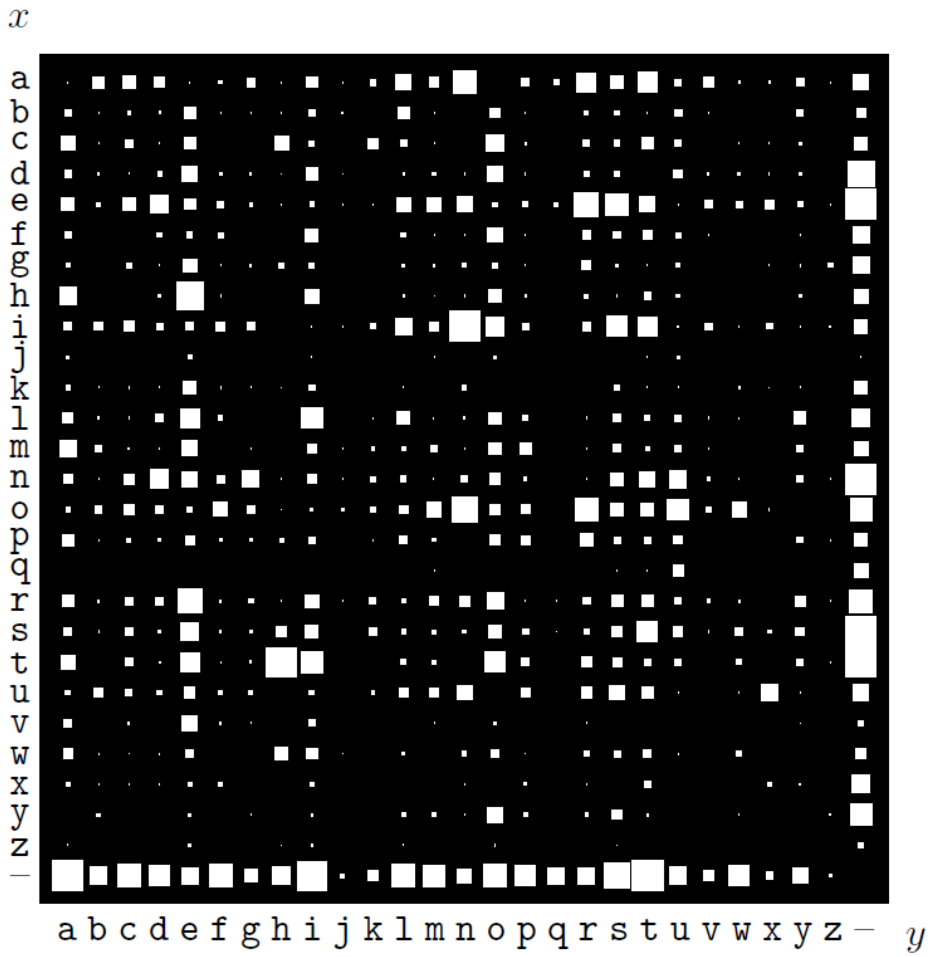
101 ?? ⟨ 101 $x_0 x_1$; 101 x_2

PREFIX CONDITION

ANY CODE CANNOT BE INITIAL PORTION OF ANOTHER CODE

EXAMPLE OF PROBABILITY DISTRIBUTIONS

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



SINGLE LETTER

TWO CONSECUTIVE LETTERS


ANDREW
DAVID
LUCY
GRAHAM
ANNE
DANIEL
RICHARD
DANIEL
GRAHAM
DAVID
AMY
JOHN
KATE
JOSHUA
LAUREN
MEGAN
PAUL
SARAH
THOMAS
SIMON
MICHAEL
KEVIN
PETER
DAVID

LABEL INVARIANCE

$$X \quad \begin{cases} x_1 & x_i & x_n \\ p_1 & p_i & p_n \end{cases} \qquad Y \quad \begin{cases} y_1 & y_i & y_n \\ p_1 & p_i & p_n \end{cases}$$

TWO RANDOM VARIABLE WITH DIFFERENT OUTCOMES BUT SAME PROBABILITY DISTRIBUTION

$$H = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$$

IS THE SAME

LABEL INVARIANCE → ENTROPY DOES NOT DEPEND ON OUTCOMES BUT ONLY ON PROB.

FUNCTION OF A RANDOM VAR.

X $\quad \Omega_X = \{ x_0 \quad x_i \quad x_n \}$

f

FUNCTION $\quad \Omega_Y = \{ f(x_0) \quad f(x_i) \quad f(x_n) \}$

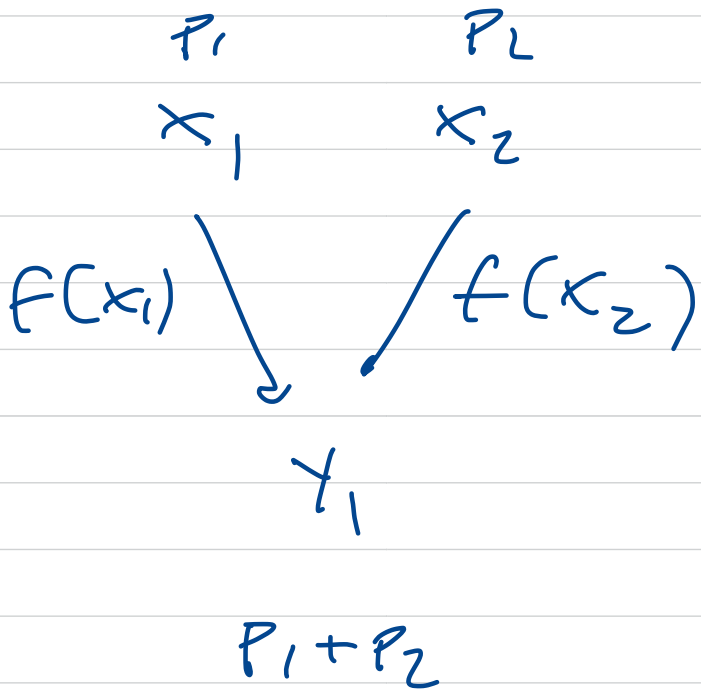
$Y = f(x)$ $\quad H(x)$ $\quad H(y)$

IF f IS ONE-TO-ONE

p_1	p_i	p_n
x_1	x_i	x_n
↓	↓	↓
$f(x_1)$	$f(x_i)$	$f(x_n)$
p_1	p_i	p_n

$$H(x) = H(y)$$

IF F IS NOT ONE-TO-ONE



$H(f(x)) \leq H(x)$

PROOF

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2}$$

$$(p_1 + p_2) \log_2 \frac{1}{p_1 + p_2}$$

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} - (p_1 + p_2) \log_2 \frac{1}{p_1 + p_2}$$

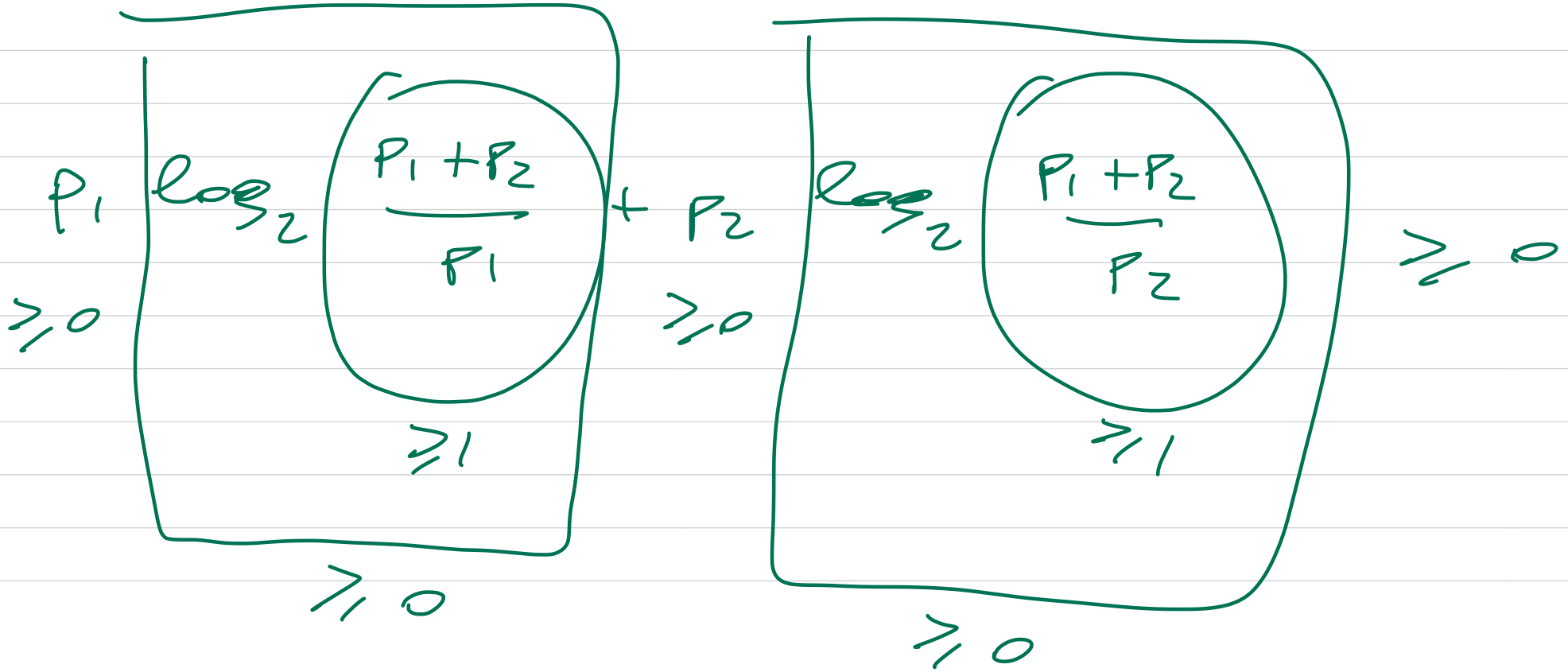
$$- p_1 \log_2 \frac{1}{p_1 + p_2}$$

$$- p_2 \log_2 \frac{1}{p_1 + p_2}$$

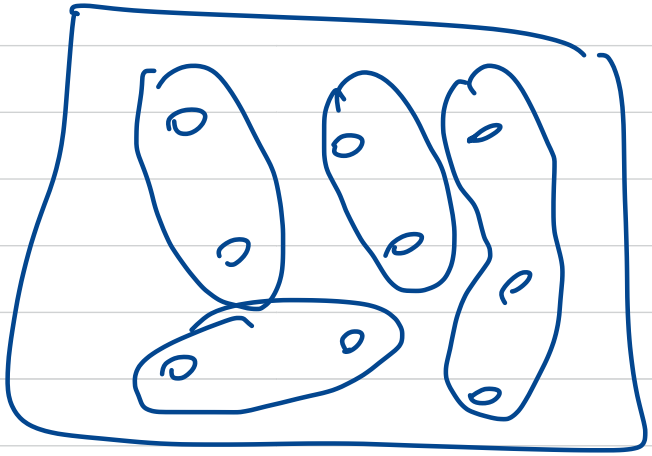
$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} - (p_1+p_2) \log_2 \frac{1}{p_1+p_2}$$

$$- p_1 \log_2 \frac{1}{p_1+p_2}$$

$$- p_2 \log_2 \frac{1}{p_1+p_2}$$



EXAMPLE



CLUSTERING IS A FUNCTION WHICH IS NOT ONE-TO-ONE

CLUSTERING ALWAYS DECREASES THE ENTROPY (I.E., THE AMOUNT OF INFORMATION)

ALTERNATIVE ENTROPY DEFINITIONS FOR ANOMALY DETECTION

DATA SERIES

RANDOM DATA OR CONTAINS SOME REGULARITY?

1 0 1 0 1 0 1 0 1 0

0 1 1 0 1 1 1 0 1 0

KOLMOGOROV COMPLEXITY OF A SEQUENCE

MINIMUM SIZE OF A MESSAGE (PROGRAM) WHICH ALLOWS TO REBUILD THE SEQUENCE

IF THE SEQUENCE IS COMPLETELY RANDOM

THE SIZE OF THE MESSAGE

IS EQUAL TO THE SIZE OF THE

SEQUENCE

PERMUTATION ENTROPY

2 8 9 7 5 2 3 5 4 1 6

ORDER $m = 3$

WE ANALYZE VECTORS OF LENGTH 3 INSIDE THE SEQUENCE

2 8 9 7 5 2 3 5 4
8 9 7 5 2 3 5 4 1
9 7 5 2 3 5 4 1 6

2 8 9 7 5 2 3 5 4
8 9 7 5 2 3 5 4 1
9 7 5 2 3 5 4 1 6

1 2 3 3 3 1 1 3 2
2 3 2 2 1 2 3 2 1
3 1 1 1 2 3 2 1 3

$m = 3$

1 1 2 2 3 3
2 3 1 3 1 2
3 2 3 1 2 1

2 1 1 1 1 3 | 9

$\frac{2}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{3}{9}$

$P_1\ P_2\ P_3\ P_4\ P_5\ P_6$

$$H = \sum_{i=1}^{6} P_i \log_2 \frac{1}{P_i}$$

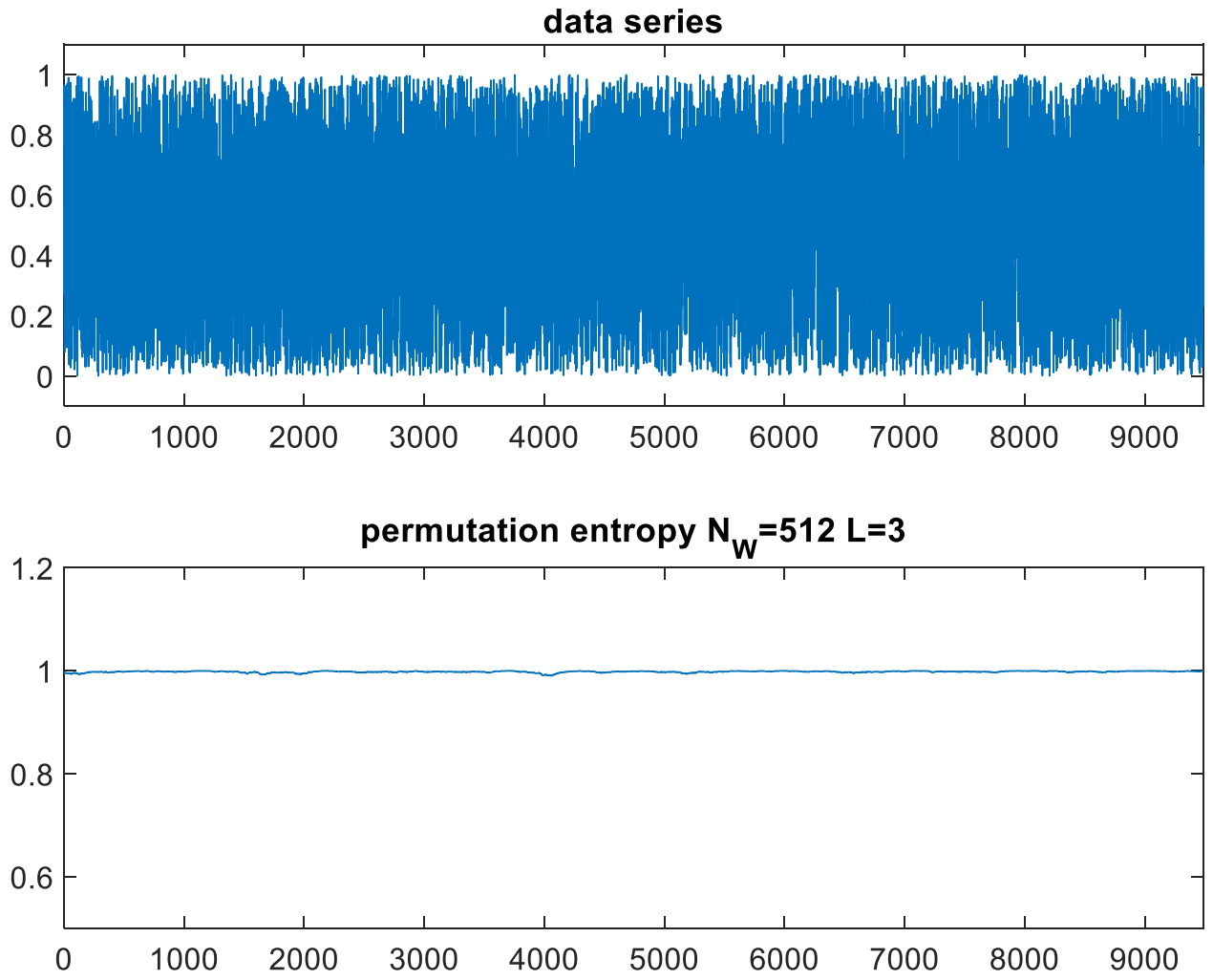
$$0 \leq H \leq \log_2 m!$$

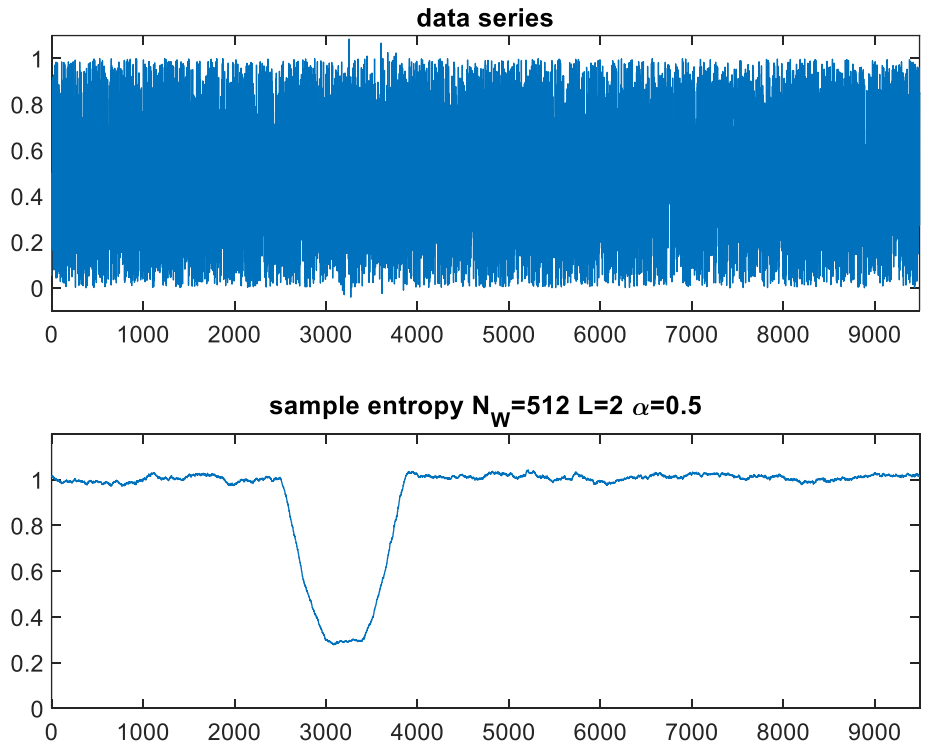
$$0 \leq \bar{H} = \frac{\sum p_i \log_2 \frac{1}{p_i}}{\log_2 m!} \leq 1$$

RANDOM DATA $\rightarrow$ $\bar{H}$ CLOSE TO 1

CORRELATED DATA $\rightarrow$ WE OBSERVE A DROP OF $\bar{H}$

EXAMPLE: RANDOM DATA




data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
sample entropy N_W=512 L=2 α=0.5
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

ALTERNATIVE: SAMPLE ENTROPY

$m = 3$

x x x x x
x x x x x
x x x x x

$\underline{A}$ $\underline{B}$ OF ORDER m

CHEBYCHEV DISTANCE

$$d(\underline{A}, \underline{B}) = \max_j |A_j - B_j|$$

WE COUNT HOW MANY TIME

$$d(\underline{A}\,\underline{B}) \leq t = r\,\sigma$$

r: PARAMETER ≤ 0.5

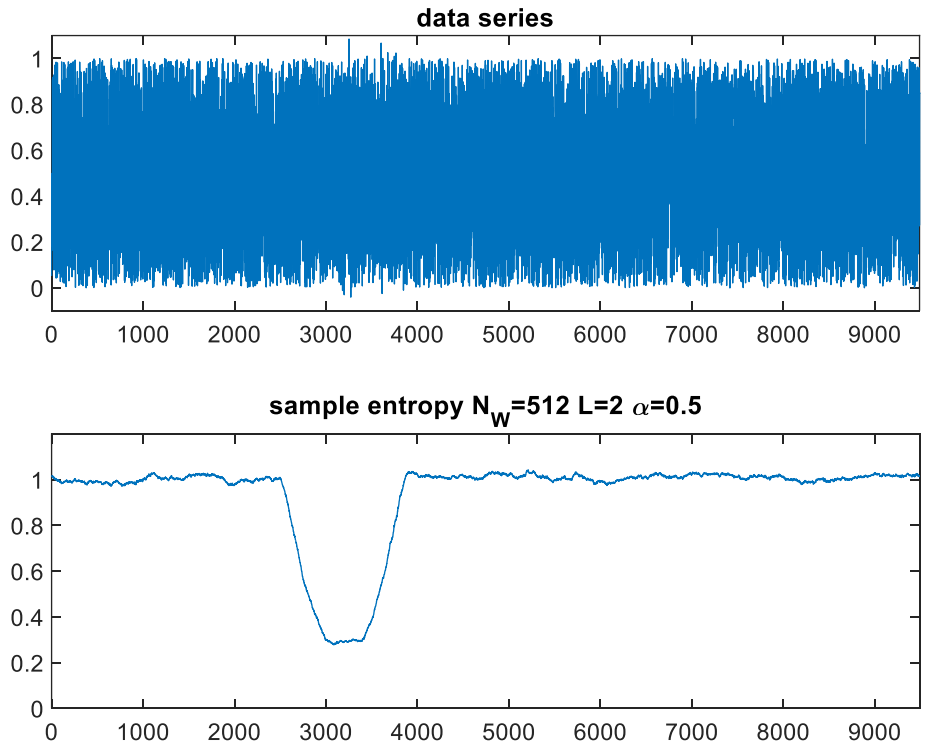
σ: St. DEV. SEQ.

C_m

WE REPEAT THIS OPERATION WITH ORDER $m+1$

C_{m+1}

SAMPLE ENTROPY $= \log_2 \frac{C_m}{C_{m+1}}$


data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
sample entropy N_W=512 L=2 α=0.5
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

JOINT ENTROPY

2 RANDOM VARIABLES

$X \quad \Omega_X$

$Y \quad \Omega_Y$

X \ Y	Temp < 25	Temp ≥ 25
Sunny	0.4	0.2
Cloudy	0.35	0.05

(X, Y)

$$H(X,Y) = \sum_{xy} p(xy) \log_2 \frac{1}{p(xy)}$$

$$0 \leq H(X,Y) \leq \log_2 \left[|\Omega_X| |\Omega_Y| \right]$$

WE WANT TO STUDY RELATIONSHIP BETWEEN $H(X,Y)$ AND $H(X)\,H(Y)$

X Y ARE STATISTICALLY INDEPENDENT

$$H(X,Y) = H(X) + H(Y)$$

FOR ALL THE OTHER CASES

$$H(X,Y) \leq H(X) + H(Y)$$

PROOF

IMPORTANT PROPERTY BASED ON MARGINALIZATION

$$\sum_y p(x,y) = p(x)$$

$$\sum_{x,y} p(x,y)\, f(x)$$

$$= \sum_x f(x) \sum_y p(x,y)$$

$$= \sum_x f(x)\, p(x)$$

S. 1. $p(x,y) = p(x)\,p(y)$

$$\sum_{x,y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

$$= \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(x)\,p(y)} =$$

$$= \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(x)}$$

$$+ \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(y)}$$

$$= \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(x)}$$

$$+ \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(y)}$$

$$= \sum_{x} p(x) \log_2 \frac{1}{p(x)}$$

$$+ \sum_{y} p(y) \log_2 \frac{1}{p(y)}$$

$$= H(X) + H(Y)$$

PROOF 2

$$\sum_{xy} p(xy) \log_2 \frac{1}{p(xy)} - \sum_x p(x) \log_2 \frac{1}{p(x)}$$

$$- \sum_y p(y) \log_2 \frac{1}{p(y)}$$

$$= \sum_{xy} p(xy) \log_2 \frac{1}{p(xy)} - \sum_{xy} p(xy) \log_2 \frac{1}{p(x)}$$

$$- \sum_{xy} p(xy) \log_2 \frac{1}{p(y)}$$

$$= \sum_{xy} p(xy) \log_2 \frac{1}{p(xy)} - \sum_{xy} p(xy) \log_2 \frac{1}{p(x)}$$

$$- \sum_{xy} p(xy) \log_2 \frac{1}{p(y)}$$

$$= \sum_{xy} p(xy) \log_2 \frac{p(x)\,p(y)}{p(xy)}$$

$$\leq \log_2 e \sum_{xy} p(xy) \left[\frac{p(x)\,p(y)}{p(xy)} - 1 \right]$$

$$\leq \log_2 e \sum_{x,y} p(x,y) \left[\frac{p(x)\,p(y)}{p(x,y)} - 1 \right]$$

$$= \log_2 e \left[\sum_{x,y} p(x)\,p(y) - \sum_{x,y} p(x,y) \right]$$

$$\sum_x p(x) \sum_y p(y)$$

$$1 \cdot 1$$

$$1 - 1 = 0$$

$$H(X,Y) - H(X) - H(Y) \leq 0$$

CONDITIONAL ENTROPY

$$H(x,y) = \sum_{xy} p(xy) \log_2 \frac{1}{p(x,y)} = (*)$$

$$p(x,y) = p(x|y)\, p(y)$$

$$(*) = \sum p(xy) \log_2 \frac{1}{p(x|y) p(y)} =$$

$$= \sum p(x,y) \log_2 \frac{1}{p(x|y)p(y)} =$$

$$= \boxed{\sum_{x,y} p(x,y) \log_2 \frac{1}{p(x|y)}} \; H(X|Y)$$

$$+ \sum_{x,y} p(x,y) \log_2 \frac{1}{p(y)}$$

$$\sum_{y} p(y) \log_2 \frac{1}{p(y)} \rightarrow H(Y)$$

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

$H(X)$ = UNCERT. ABOUT X

$$H(X|Y) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x|y)}$$

IMPORTANT

$H(X|Y)$ IS THE RESIDUAL UNCERT. ABOUT X WHEN THE OUTCOME OF Y IS REVEALED

IF X AND Y ARE ST. INDEP.

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

EVEN IF WE OBSERVE Y

WE GAIN NO INFO ABOUT X

PROOF

$$H(x,y) = H(x|y) + H(y)$$

IF x AND y ARE S.I.

$$H(x,y) = H(x) + H(y)$$

$$H(x) + H(y) = H(x|y) + H(y)$$

$$0 \leq H(X|Y) \leq H(X)$$

PROOF

IF X AND Y NOT S.I,

$$H(XY) \leq H(X) + H(Y)$$

$$H(XY) = H(X|Y) + H(Y)$$

$$H(X|Y) \leq H(X)$$

$$0 \leq H(X|Y) \leq H(X)$$

IF Y CONTAINS ALL INFO ABOUT X $H(X|Y) = 0$

(example: Y IS A ONE-TO-ONE FUNCTION OF X)

CHAIN RULE OF ENTROPY

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

$$H(XYZ) = H(X|YZ) + H(Y|Z) + H(Z)$$

IMPORTANT: INFORMATION GAIN

$$H(x)$$

$$H(x|y)$$

$$I(x,y) = H(x) - H(x|y)$$

HOW MUCH WE GAIN ABOUT x KNOWLEDGE BY OBSERVING y OUTCOME

$$0 \leq I(x,y) \leq H(x)$$

① $I(X,Y) = H(X) - H(X|Y)$

$H(XY) = H(X|Y) + H(Y)$

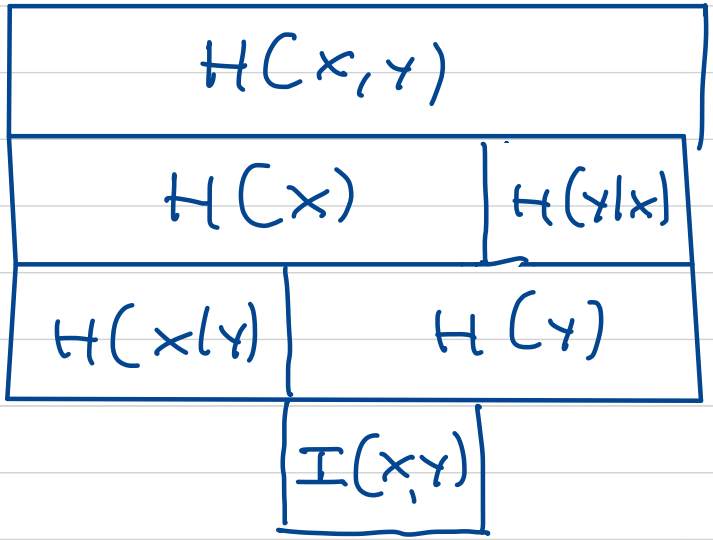
② $I(XY) = H(X) + H(Y) - H(X,Y)$

$H(YX) = H(Y|X) + H(X)$

③ $I(XY) = H(Y) - H(Y|X)$

INF. GAIN IS SYMMETRIC

IMPORTANT



$$H(xy) \leq H(x) + H(y)$$

$$H(x,y) = H(y|x) + H(x)$$

$$H(x,y) = H(x|y) + H(y)$$

$$I(xy) = H(x) - H(x|y)$$

$$I(xy) = H(y) - H(y|x)$$

$$I(xy) = H(x) + H(y) - H(xy)$$

EXAMPLE 1

P(X,Y) =
0.400 0.200
0.350 0.050

P(X) =
0.600 0.400

P(Y) =
0.750 0.250

H(X,Y) = 1.739 H(X) = 0.971 H (X|Y) = 0.928 I(X;Y) = 0.043
H(X,Y) = 1.739 H(Y) = 0.811 H (Y|X) = 0.768 I(Y;X) = 0.043

pmi =
-0.170 0.415
0.222 -1.000

E_Z = 0.043

EXAMPLE 2

```
P(X,Y) =
0.700 0.020
0.080 0.200

P(X) =
0.720 0.280

P(Y) =
0.780 0.220

H(X,Y) = 1.229    H(X) = 0.855    H (X|Y) = 0.469    I(X;Y) = 0.387
H(X,Y) = 1.229    H(Y) = 0.760    H (Y|X) = 0.374    I(Y;X) = 0.387

pmi =
0.318 -2.986
-1.449 1.699

E_Z = 0.387
```

KULLBACK-LEIBLER DIVERGENCE

$$\Omega = \{ x_0 \quad x_i \quad x_n \}$$

$$P = \{ p_1 \quad p_i \quad p_n \}$$ ← OBSERVED DISTRIBUTION

$$Q = \{ q_1 \quad q_i \quad q_n \}$$ ← DISTRIBUTION MODEL

$$D_{KL}(P|Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

IF $p_i = q_i \quad \forall i \qquad D(P|Q) = 0$

PROPERTIES SIMILAR TO DISTANCE

BUT NOT A DISTANCE

$$D(P|Q) \neq D(Q|P)$$

$$D_{KL}(P|Q) \geq 0$$

PROOF

$$D = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

$$-D = \sum_i p_i \log_2 \frac{q_i}{p_i}$$

$$\leq \log_2 e \sum_i p_i \left(\frac{q_i}{p_i} - 1 \right)$$

$$= \log_2 e \left[\sum_i q_i - \sum_i p_i \right] = 0$$

$$1 \qquad 1$$

$$-D \leq 0 \rightarrow D \geq 0$$

LINK BETWEEN D AND INF. GAIN

$$I(x,y) = H(x) + H(y) - H(x,y)$$

$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)}$$

$$= D_{KL}\left(p(x,y) \,|\, p(x)\,p(y) \right)$$

$$D_{KL}(P|Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

P OBSERVED

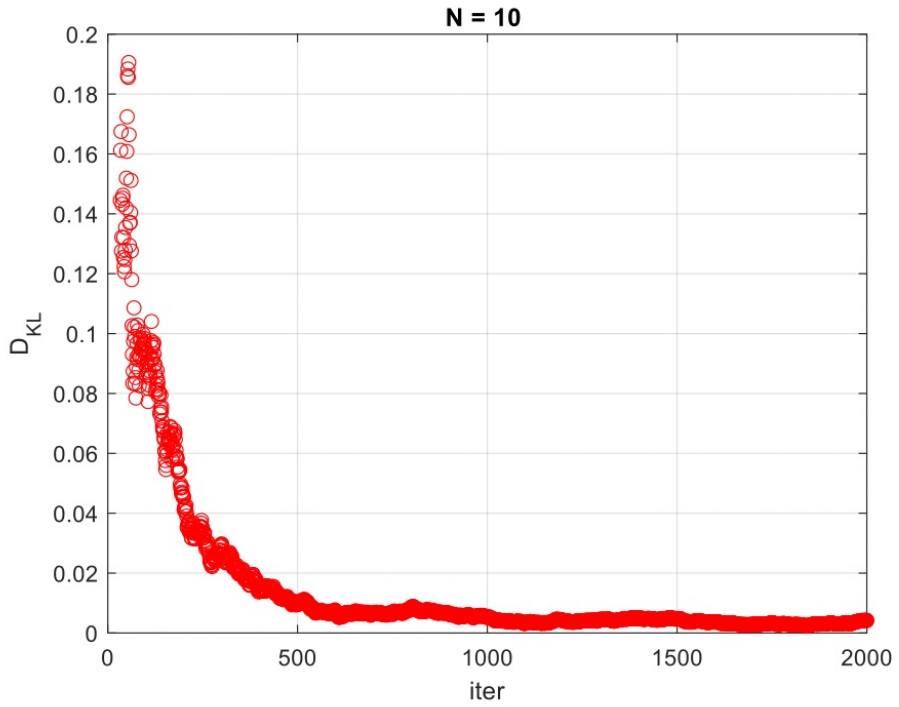
Q MODEL

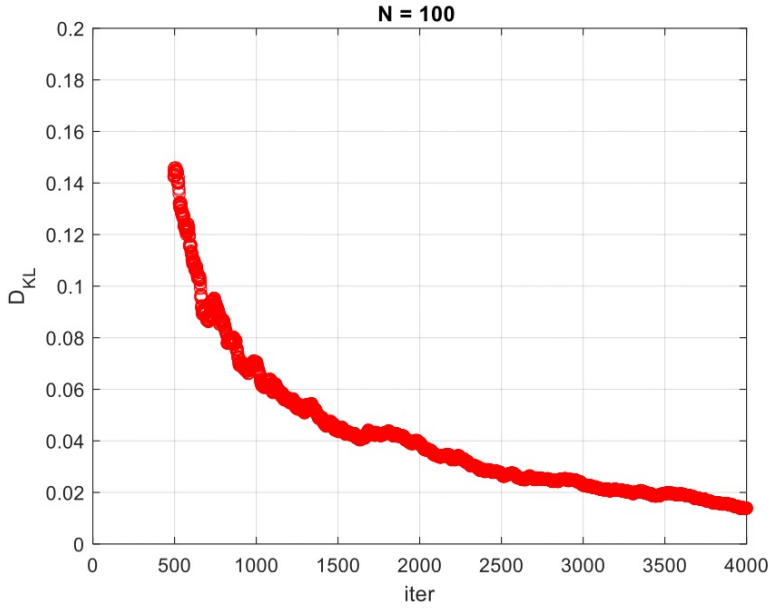
D_{KL} REPRESENTS THE INFO. GAIN WHEN WE USE THE OBSERVED P INSTEAD OF THE MODEL DISTR. Q.

$\equiv$ THE INFO LOSS WHEN WE USE MODEL Q INSTEAD OF OBSERVED P

EXAMPLE 1: OBSERVED DATA = RANDOMLY GENERATED WITH UNIFORM DISTRIBUTION BETWEEN 1 AND N

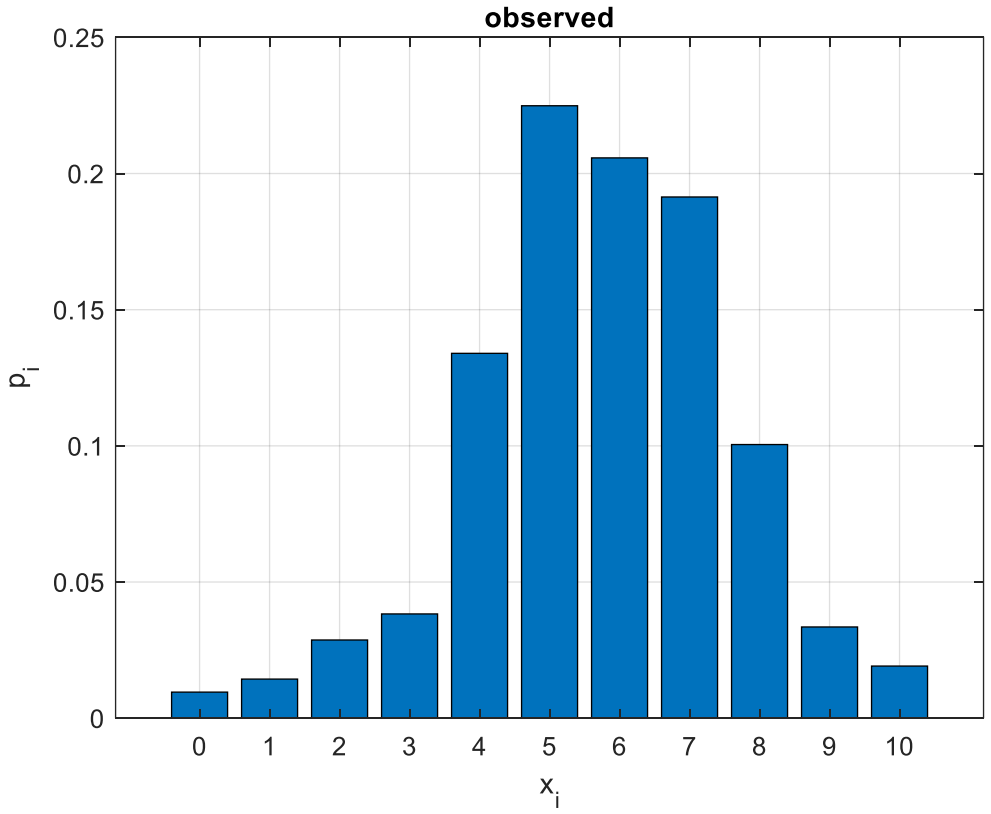
VS. IDEAL UNIFORM DISTRIBUTION BETWEEN 1 AND N




N = 100
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
D_{KL}
0
500
1000
1500
2000
2500
3000
3500
4000
iter

EXAMPLE 2

OBSERVED DISTRIBUTION
WE LOOK FOR THE BEST FIT
WITH BINOMIAL DISTRIBUTION



Information Theory for Data Science

Assignment 1

Introduction to Information Theory and application to Classifiers

Draft version 0.2

Exercises:

1. Entropy of a binary random variable with 3 outcomes (pt. X)

2. Application of the principle of maximum entropy (pt. X)

3. Permutation Entropy for Time Series Anomaly Detection (pt. X)

4. Kullback Leibler divergence from empirical distribution (pt. X)

Exercise 1 - Entropy of a binary random variable

Exercise 1.A

1. Plot the entropy of a binary random variable
2. Discuss the result

Exercise 1.B – Renyi entropy

$$H_\alpha(X) = \frac{1}{1-\alpha}\log_2\left[\sum_i p_i^\alpha\right] \qquad \alpha \geq 0 \qquad \alpha \neq 1$$

3. Prove that the limit for $\alpha \to 1$ is the Shannon entropy.

4. Repeat point 1 with different values of α (smaller and bigger than 1)

5. Comment the results

Exercise 1.C

- Prove that this averaging operation $\{p, p, p_3\}$ $p = \frac{p_1 + p_2}{2}$ always increases the entropy

Hint: use the log inequality

Exercise 2 - Application of the principle of maximum entropy

Exercise 2

1. Invent an exercise where you have a random variable X with alphabet Ω_X where each outcome has a given “cost”.
2. Fix the mean value bigger than the arithmetic average of the costs, and apply the principle of maximum entropy to find the probability distribution P(X)
3. Plot P(X)
4. Repeat with a mean value equal to the arithmetic average and plot the result
5. Repeat with other values of the mean value and plot the results
6. Comment the results

You must numerically solve the equation generated by the Lagrange optimization.

As an example , for Matlab you can use

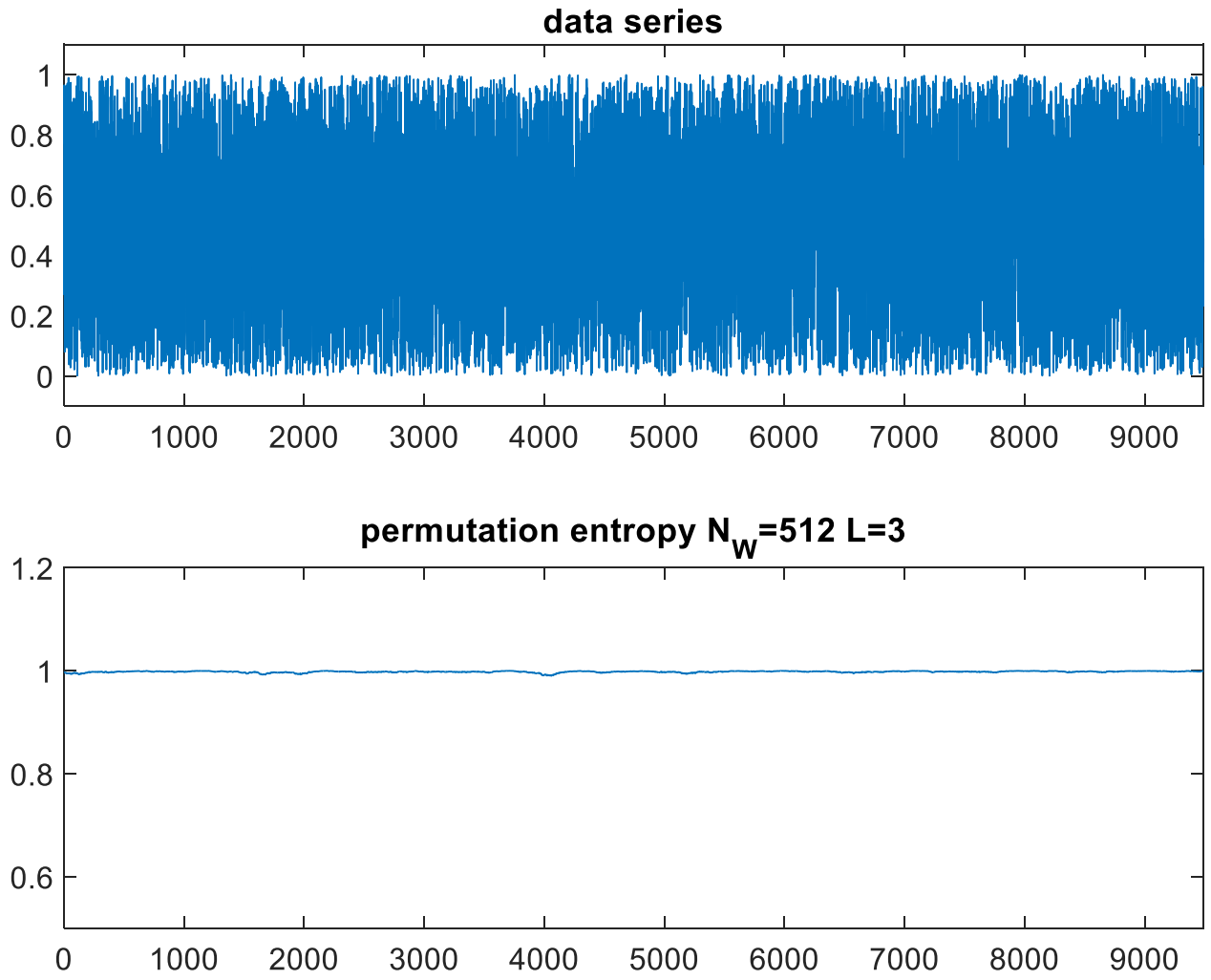
```
syms x
eqn = ( . . .  )*mu == ( . . . );
V = vpasolve(eqn,x,[0 10])
```

Exercise 3 - Permutation Entropy for Time Series Anomaly Detection

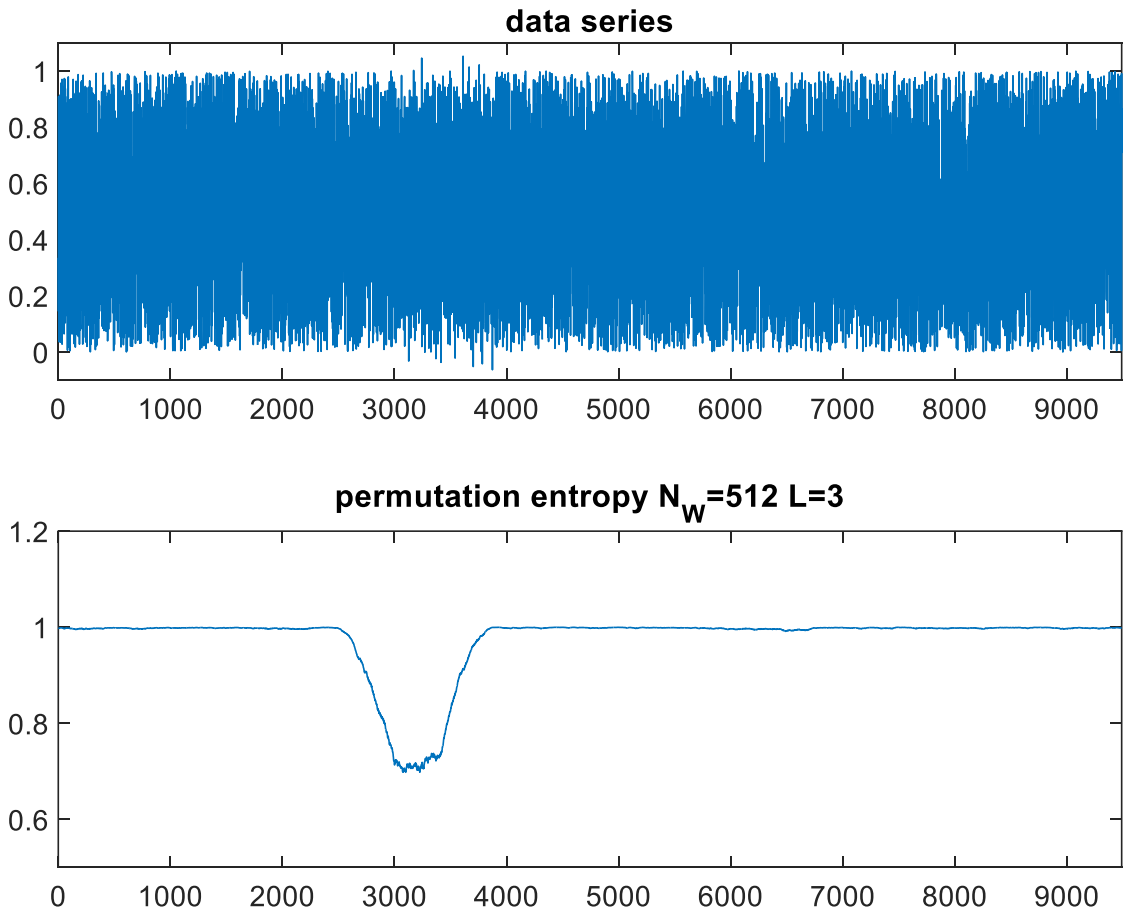
Exercise 3

Write a program that:

- Generates a data series made by 10,000 random symbols.
- Computes the permutation entropy with a sliding window of $N_W = 512$ symbols. (For example, choose an order m = 3.)
- Plot the data and the entropy.


data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
permutation entropy N_W=512 L=3
1.2
1
0.8
0.6
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

- Insert a pattern between 3000 and 4000 made by correlated data (with about the same mean value and variance of the original random data).
- Describe how you generated the correlated data
- Compute the permutation entropy with the same N_W and N_R used before.
- Plot the data and the entropy.
- Comment the results


data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
permutation entropy N_W=512 L=3
1.2
1
0.8
0.6
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

Some Matlab functions that might be useful

```
sort
perms
```

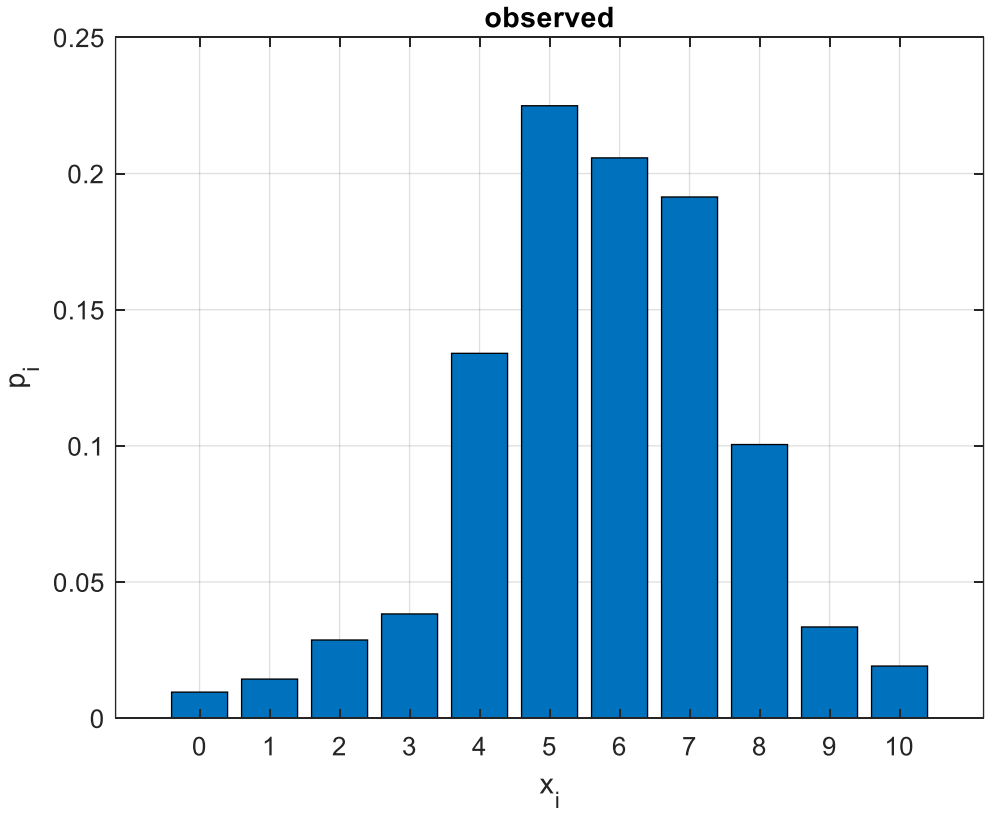
Exercise 4 - Kullback-Leibler distance from empirical distribution

Exercise 4

Consider the observed data

$$x_i = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

$$\mathrm{NUM}\left(x_i\right) = \begin{bmatrix} 1 & 3 & 2 & 8 & 22 & 45 & 44 & 42 & 24 & 8 & 3 \end{bmatrix}$$


observed
0.25
0.2
0.15
0.1
0.05
0
p_i
0
1
2
3
4
5
6
7
8
9
10
x_i

1. Compare with uniform pmf
2. Write in the title the KL divergence value

Compare with binomial pmf with $0 < p < 1$ (step=0.001)

3. Identify the pmf at minimum KL divergence
4. Plot the pmf
5. In the title write the value of p and KL divergence