Skoltech

Skolkovo Institute of Science and Technology

Stochastic methods in Mathematical Modelling

Lecture 6. Information-Theoretic View on Randomness

Skoltech
Skolkovo Institute of Science and Technology

Information content

What is an information content of a random outcome?

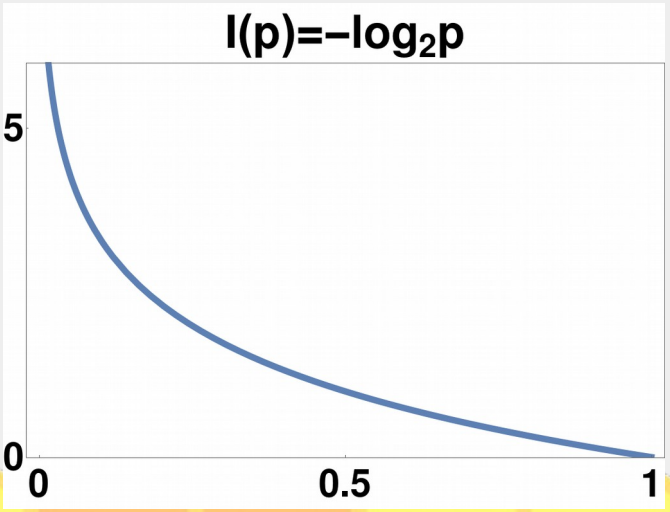
Considerations (Shannon, 1948)

1) If something happens with certainty the information content (self-information) is zero
2) Less probable events give more information
3) For two independent events the total information content should be sum of two self-informations

$$I(x) \equiv -\log_b P(x)$$

If $b = 2$ then the information content is measured in *bits.*
Alternatively, b = e could be used (*nats*)
as well as some other choices

Intuitively, $I(x)$ is a measure of “surprise”



Skoltech
Skolkovo Institute of Science and Technology

Information entropy

What is an information content of a random outcome?

The *expectation value* of the self-information is called the *entropy* of a random variable

$$H(x) \equiv E[I(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Properties

1) $H(X) \geq 0$
2) $H(X) = 0$ iff X is deterministic
3) $H(X) \leq \log_2 n$. The equality $H(x)=\log_2 n$ is achieved iff all outcomes are equiprobable
4) Entropy is less or equal than the average number of bits required to describe the random variable

Jensen’s inequality and proof of 3rd property

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Intuitive way of obtaining the formula for information entropy

Let's construct the message from N independent outcomes of a random variable with M possible outcomes for every symbol. In order to convey the message precisely one needs $N \log_2 M$ bits of information.

The probabilities $\{p_i\}$ constrain the likely types of messages.

One expects that for large N a message contains $N_i = Np_i$ happenings of a symbol i

The typical message is then a number of ways of rearranging N_i occurrences of i. That is

$$g = \frac{N!}{\prod_{i=1}^{M} N_i!}$$

Less than total M^N

Specifying one out of g sequences requires $\log_2 g \approx -N \sum_{i=1}^{M} p_i \log_2 p_i$ bits

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

What is an information content of a random outcome?

The expected value of the self-information is called the entropy of a random variable

$$H(x) \equiv E[I(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Properties

1) $H(X) \geq 0$
2) $H(X)=0$ iff X is deterministic
3) $H(X) \leq \log_2 n$. The equality $H(X)=\log_2 n$ is achieved Iff all outcomes are equiprobable
4) The choice of the logarithm base is custom. It just rescales the quantity
5) Entropy is a lower bound on the average length of the shortest description of random variable

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Exercise:

Compare three cases in terms of entropy

(a) 5 equally probable states;
(b) 3 states which happens with the probabilities 1/2, 1/6, 1/3;
(c) 6 states which happen with the probabilities 1/2, 1/10, 1/10, 1/10, 1/10, 1/10.

$$H(x) \equiv E[I(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

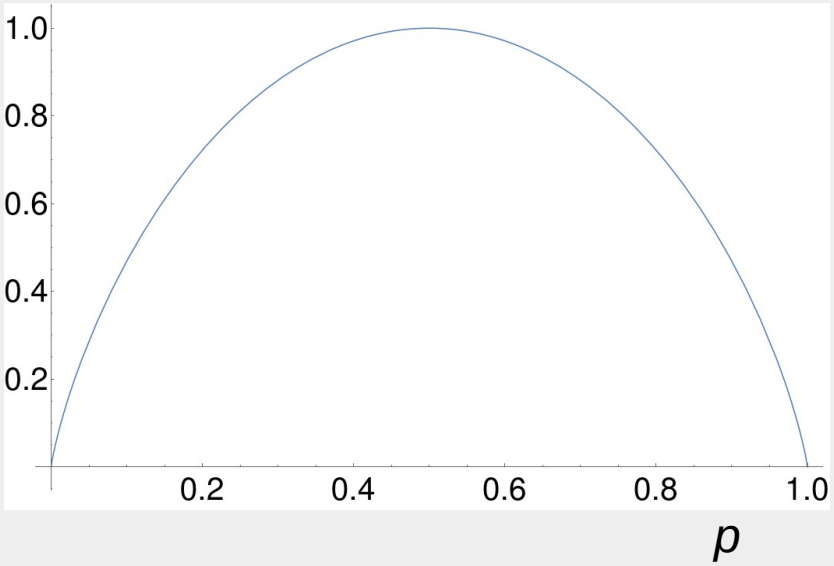
Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Example: Bernoulli distribution

$$H_{binary}(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

$H_{binary}(p)$



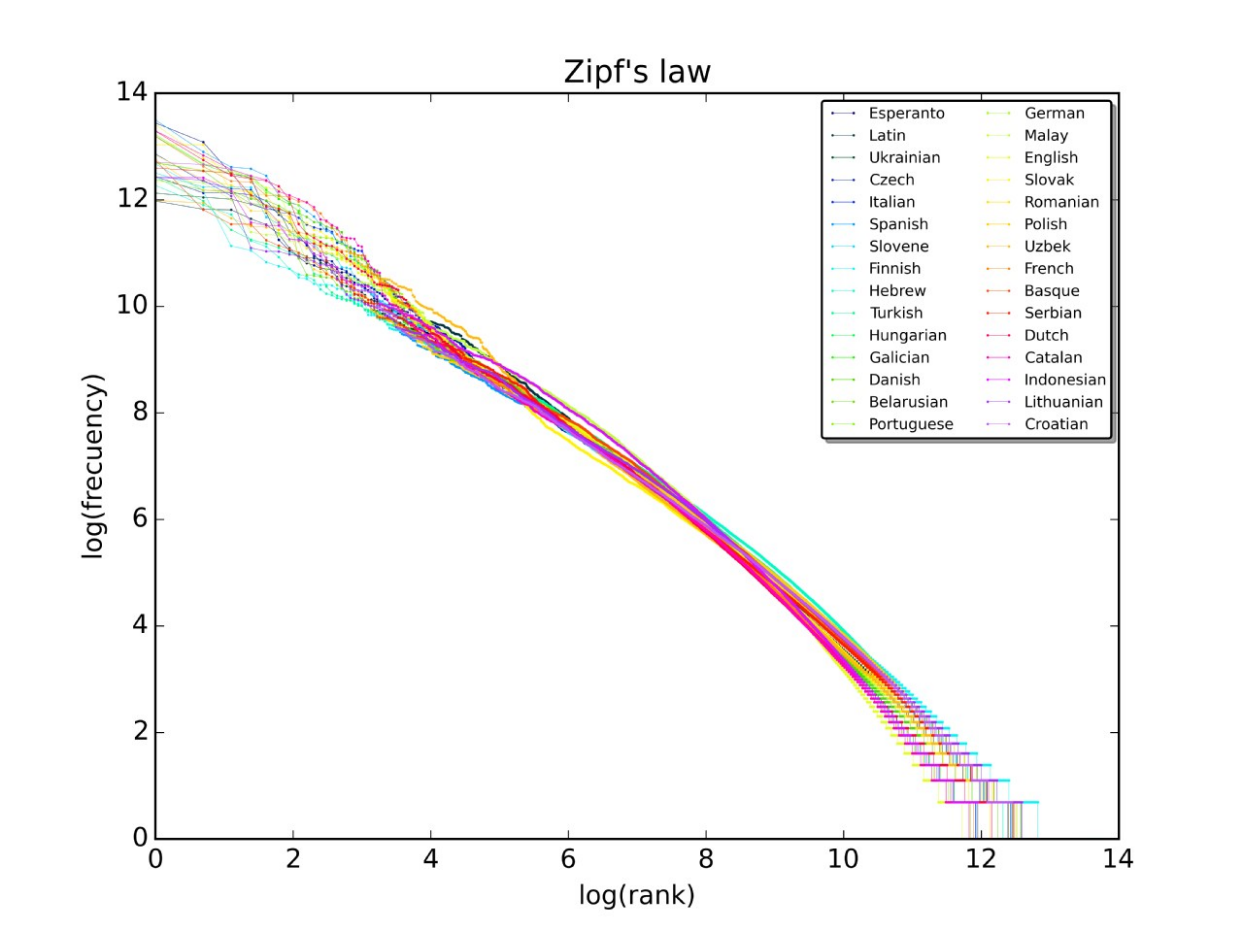
1 bit of information is required to define the maximum

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Example: Entropy of English language&Zipf's law $1/n^{\alpha}$

frequency of a word is inversely proportional to its rank in the frequency table



The most frequent words of English:

1. the
2. be
3. to
4. of
5. and

…

55. time

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a log-log scale

taken from https://en.wikipedia.org/wiki/Zipf%27s_law

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Example: Entropy of “English” language&Zipf’s law

frequency of a word is inversely proportional to its rank in the frequency table

$$p_n = \begin{cases} \frac{0.1}{n}, for\ n = 1,2,\dots,12367 \\ 0, for\ n > 12367 \end{cases}$$

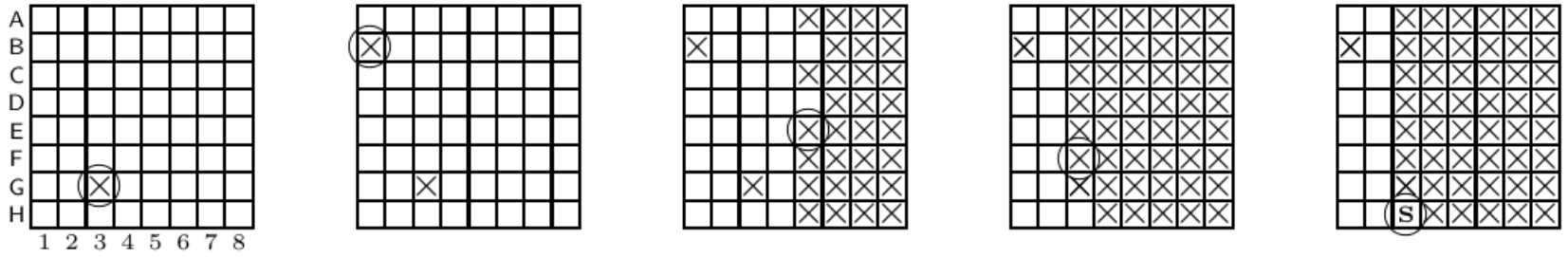
Compute the entropy of this made-up English language per word

$$H(x) \equiv E[I(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Skoltech
Skolkovo Institute of Science and Technology

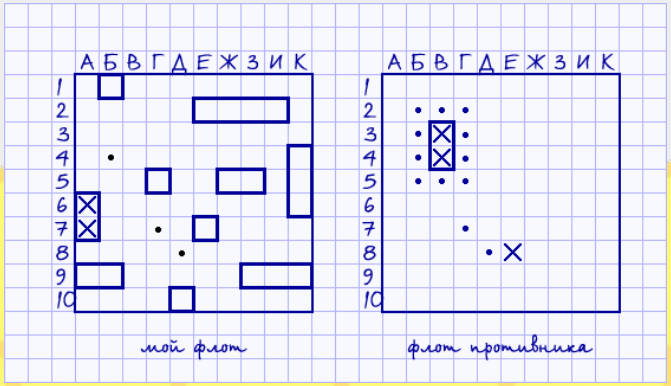
Shannon Information

Example: The game of submarine



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

The game of battleships



Skoltech
Skolkovo Institute of Science and Technology

How many bits are needed to describe an outcome of an experiment?

If the data can be compressed into a file with ***L*** bits per source symbol and the data can be recovered reliably then we say that the information content is at most ***L*** bits per symbol.

Example: compression of text files

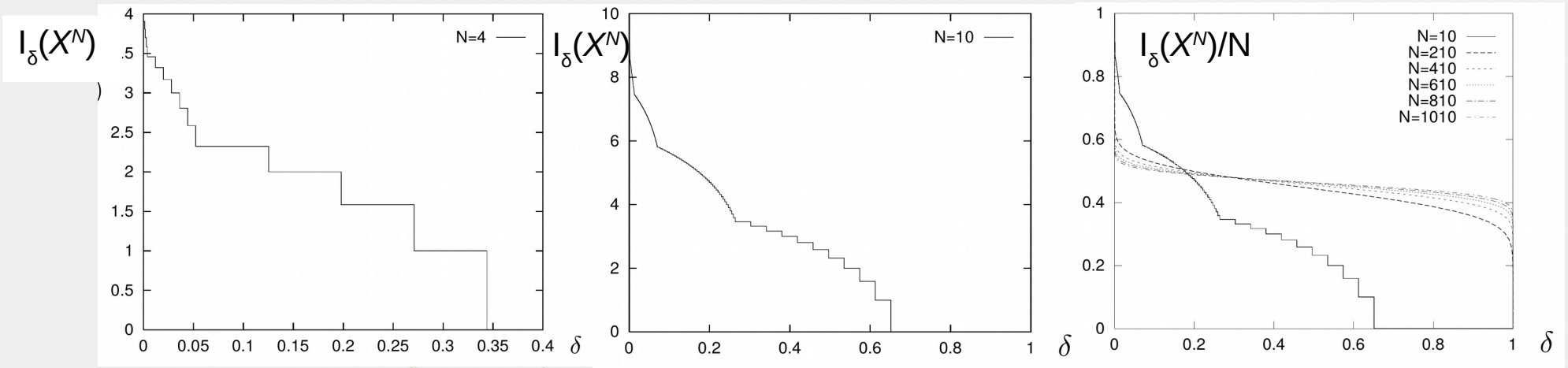
Skoltech
Skolkovo Institute of Science and Technology

Bent coin example

N throws of a bent coin **x**={x_1,x_2, …, x_N}, $x_i \in \{0,1\}$ with probabilities p_0= 0.9 and p_1= 0.1, i.e. if $r(x)$ is the number of ones in **x** then

$$P(x) = p_0^{N-r(x)}\, p_1^{r(x)}$$

To evaluate $I_\delta(X^N)$ we must find the smallest sufficient subset S_δ. This subset will contain all ***x*** with $r(\boldsymbol{x})$ = 0, 1, 2, . . . , up to some $r_{max}(\delta) - 1$, and some of the ***x*** with $r(\mathbf{x}) = r_{max}(\delta)$



Skoltech

Skolkovo Institute of Science and Technology

Shannon's source coding theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left|\frac{1}{N} H_\delta\left(X^N\right) - H\right| < \epsilon$$

N i.i.d. random variables each with entropy *H*(X) can be compressed into more than *NH*(X) bits with negligible risk of information loss, as $N \rightarrow \infty$; conversely if they are compressed into fewer than *NH*(X) bits it is virtually certain that information will be lost!

One cannot compress the data such that the average number of bits per symbol is less than the Shannon entropy of the source

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Joint entropy

$$H(X,Y) = -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} P(x_i, y_i)\log_2 P(x_i, y_i)$$

Conditional entropy

$$H(Y|X) = -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} P(x_i, y_j)\log_2 P(y_i|x_i)$$ Note: $H(Y|X) \neq H(X|Y)$

Marginal entropy

The chain rule: $H(X,Y) = H(X) + H(Y|X)$

For *n* different variables X_1, …, X_n:

$$H(X_{1,\dots}, X_n) = \sum_{i=1}^{n} H(X_i|X_i - 1,\dots, X_1)$$

Skoltech
Skolkovo Institute of Science and Technology

Information entropy

Mutual Information

$$I(X,Y)=\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} P(x_i,y_i)\log_2\frac{P(x_i,y_i)}{P(x_i)P(y_j)}$$

Measures the information X and Y share.
How much knowing one of the variables reduces the uncertainty about the other

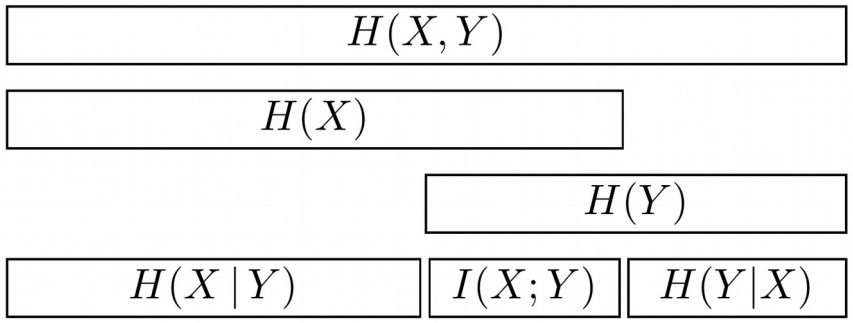
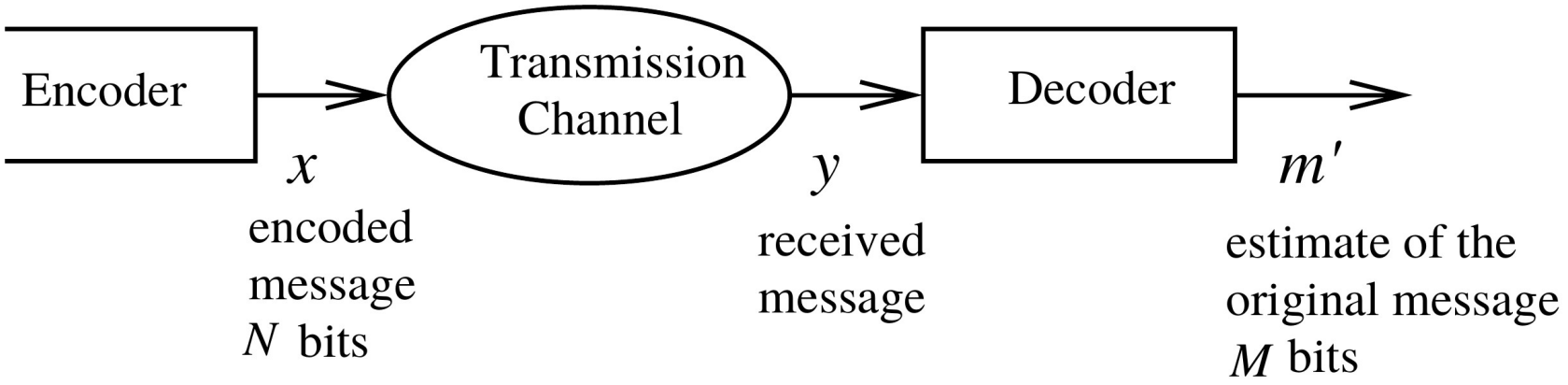


Figure 8.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy.

Skoltech
Skolkovo Institute of Science and Technology

Communications over a noisy channel



Skoltech
Skolkovo Institute of Science and Technology

Communications over a noisy channel

Binary symmetric channel

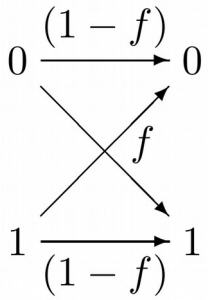
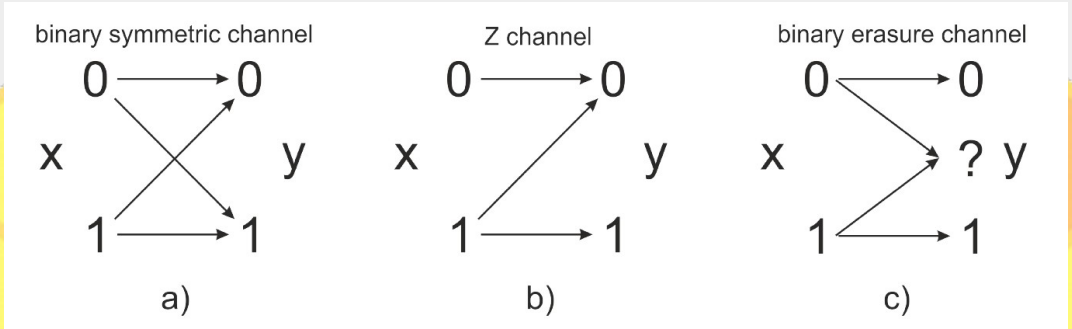


Figure 1.5. A binary data sequence of length 10 000 transmitted over a binary symmetric channel with noise level $f = 0.1$. [Dilbert image Copyright©1997 United Feature Syndicate, Inc., used with permission.]

$$x \;\begin{array}{c} 0 \to 0 \\ 1 \to 1 \end{array}\; y \qquad \begin{aligned} P(y=0\,|\,x=0) &= 1-f; & P(y=0\,|\,x=1) &= f; \\ P(y=1\,|\,x=0) &= f; & P(y=1\,|\,x=1) &= 1-f \end{aligned}$$

Examples of other standard model channels



Skoltech
Skolkovo Institute of Science and Technology

Communications over a noisy channel

Channel capacity

It appears that the mutual information between the input and the output is a function of input ensemble. Hence, maximisation of the mutual information can be done by the best possible input ensemble

Channel capacity $$C(Q) = \max_{P_X} I(X;Y)$$

P_X which optimises the capacity is “optimal input distribution”

Capacity measures the maximum rate of error-free information that can be transmitted through the channel

Skoltech
Skolkovo Institute of Science and Technology

Communications over a noisy channel

Binary symmetric channel exercise

Consider a BSC with the error probability, $f = 0.15$, and the following input probability distribution: $P(x = 0) = 0.9$, $P(x = 1) = 0.1$. In other words, the input signal is a Bernoulli process with $p = 0.1$.

1) Calculate the output probability distribution, P(y).

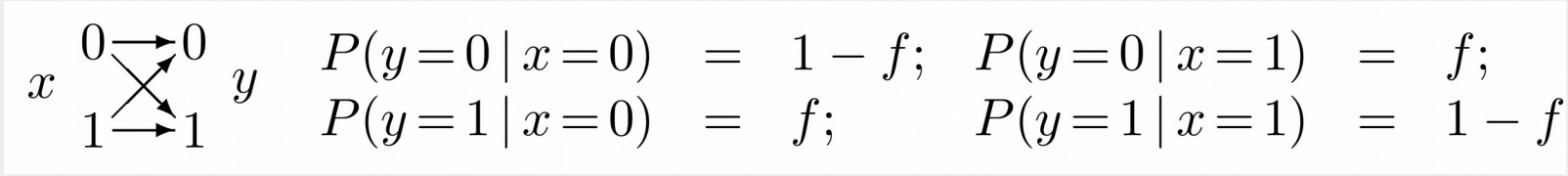
2) Compute the probability $x = 1$ given $y = 1$.

$$I(X,Y) = \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_i) \log_2 \frac{P(x_i, y_i)}{P(x_i) P(y_j)}$$

3) Compute the mutual information I(X; Y).

4) What is the capacity of the channel C(Q) as a function of *f*?

$$C(Q) = \max_{P_X} I(X;Y)$$



Skoltech
Skolkovo Institute of Science and Technology

Kullback-Leibler divergence (KL distance, KL divergence)

How does one measure a “distance” between two distributions with the same support?

$$D_{KL}(p(x)||q(x)) = \sum p(x) \log_2 \frac{p(x)}{q(x)}$$

$$D_{KL}(q(x)||p(x)) = \sum q(x) \log_2 \frac{q(x)}{p(x)}$$

non-symmetric measure of the difference between two probability distributions p(x) and q(x). Not a proper distance!!

$$D_{KL}(q(x)||p(x)) \neq D_{KL}(p(x)||q(x))$$

$$D_{KL}(q(x)||p(x)) \geq 0 ; D_{KL}(p(x)||q(x)) = 0 \; iff \; P = Q$$

$$D_{KL}(p(x)||q(x)) = -H(P) - \sum P \log Q$$

cross entropy

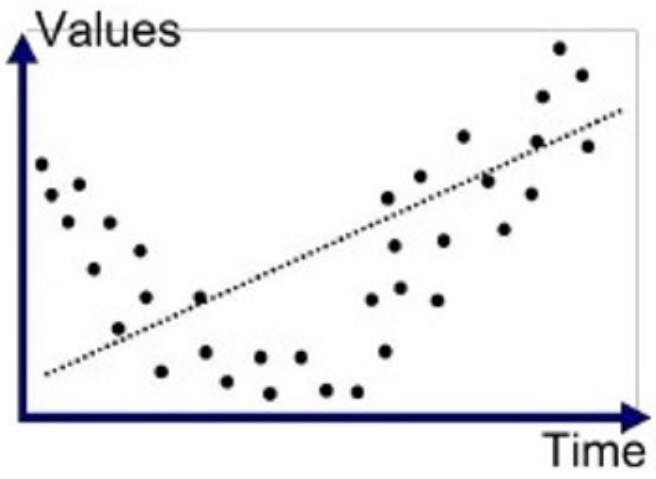
Skoltech

Skolkovo Institute of Science and Technology

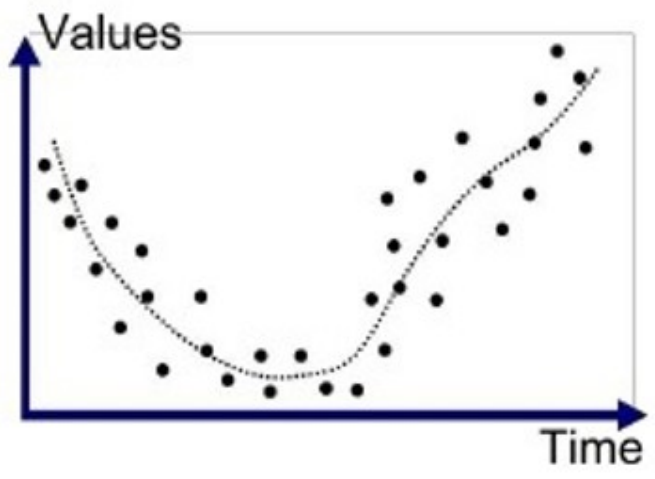
Information criteria (AIC,BIC,TIC etc.)

How do we understand which model is better?

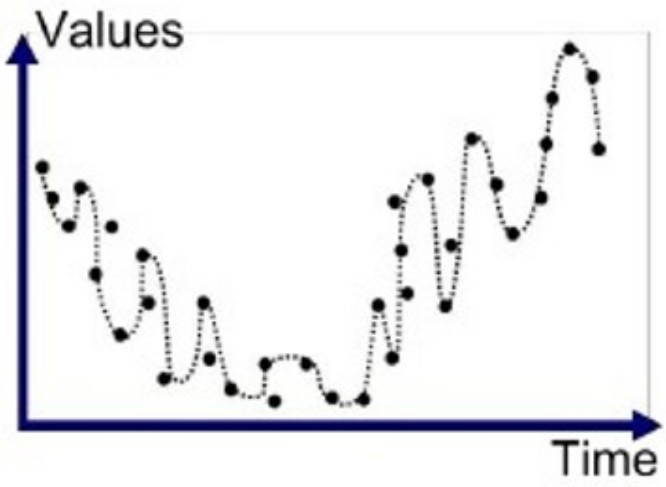
More parameters and better fit or otherwise?



Underfitted



Good Fit/Robust



Overfitted

Skoltech
Skolkovo Institute of Science and Technology

Information criteria (AIC,BIC,TIC etc.)

Means of model selection. Estimate which of the candidate models describes data in the best possible way

Most frequently used

Akaike information criterion (AIC)

$$AIC = 2d - 2\ln\left(q\left(x|\widetilde{\theta}(x)\right)\right)$$

the number of params

Bayesian information criterion (BIC)

$$BIC = d\ln n - 2\ln\left(p\left(x|\widetilde{\theta}\right)\right)$$

Sample size

According to AIC if three models have AIC values AIC_1=10, AIC_2=12, AIC_3=15 then the second model is exp((10-12)/2)=0.368 more likely than the first one and the third one is exp((10-15)/2)=0.082 more likely than the first one

Skoltech

Skolkovo Institute of Science and Technology

Information-Theoretic View on Randomness

Literature

1. D. J. C. Mackay, Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003. http://www.inference.org.uk/itprnn/book.html

2. M. Mézard, A. Montanari, Information, Physics and Computation. Oxford Univesity Press, 2009.

3. All of Statistics, A Concise Course in Statistical Inference, L. Wasserman. Springer, 2004.