



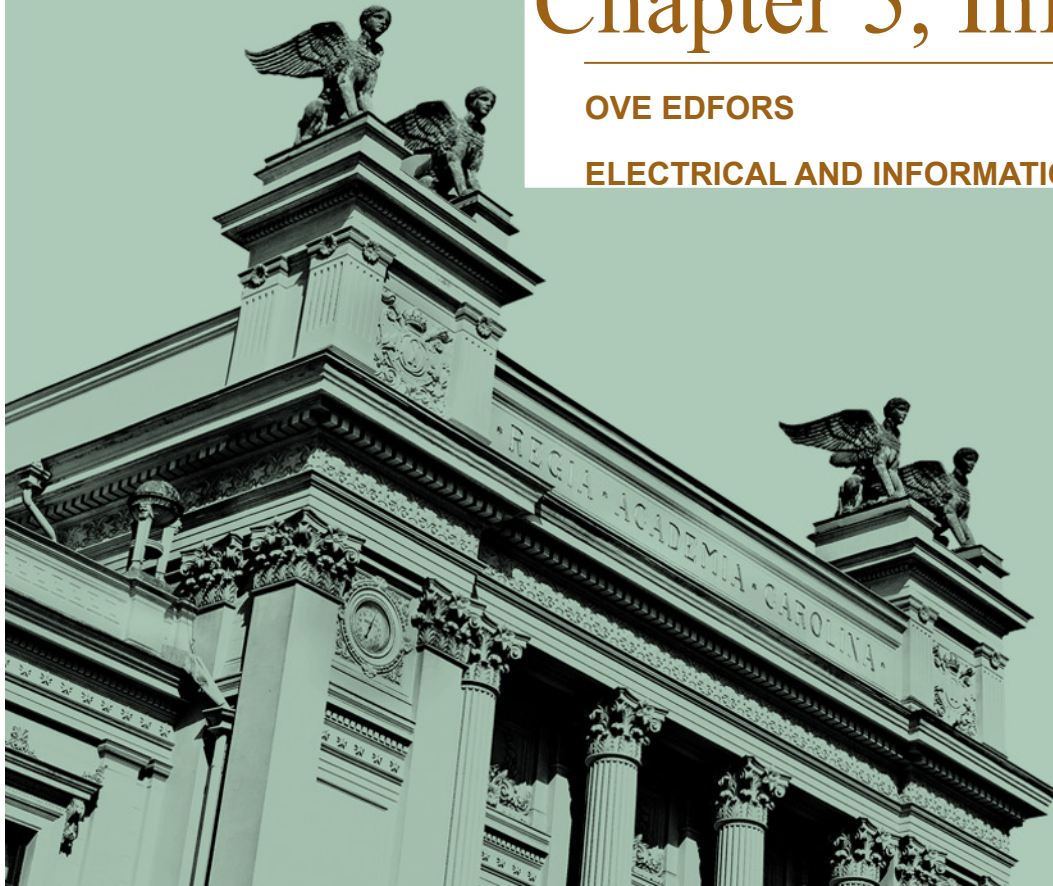
LUND
UNIVERSITY

Information Transmission

Chapter 5, Information theory

OVE EDFORS

ELECTRICAL AND INFORMATION TECHNOLOGY



Learning outcomes

- After this lecture, the student should
 - understand the mathematical concepts of
 - uncertainty, aka entropy,
 - conditional uncertainty, aka conditional entropy, and how they are calculated and some of the basic bounds.
 - understand what mutual information is and how it is calculated,
 - have a basic understanding of typical sequences and their connection to the uncertainty (entropy) of entire sequences, e.g. text, and what implications this has on source coding.



What did Shannon contribute with?



- Founded information theory in 1948 "A mathematical theory of communication"
- "one of the most important master's theses ever written": A symbolic analysis of relay and switching circuits
- Put cryptology into a mathematical framework 1949 "Communication theory of secrecy systems"



Entropy, "uncertainty"

Shannon defined the *uncertainty* or *entropy* of a discrete random variable X to be the quantity

$$H(X) \stackrel{\text{def}}{=} - \sum_{i=1}^L P_X(x_i) \log P_X(x_i)$$

The unit of the uncertainty is called *bit*.

One bit is the uncertainty of a binary random variable that is 0 and 1 with equal probability.



Uncertainty, upper and lower bound

The uncertainty $H(X)$ of the discrete random variable X with L outcomes is lower- and upper-bounded by

$$0 \leq H(X) \leq \log L$$

with equality on the left if and only if $P_X(x) = 1$ for some x , and with equality on the right if and only if $P_X(x) = 1/L$ for all x .



The binary entropy function

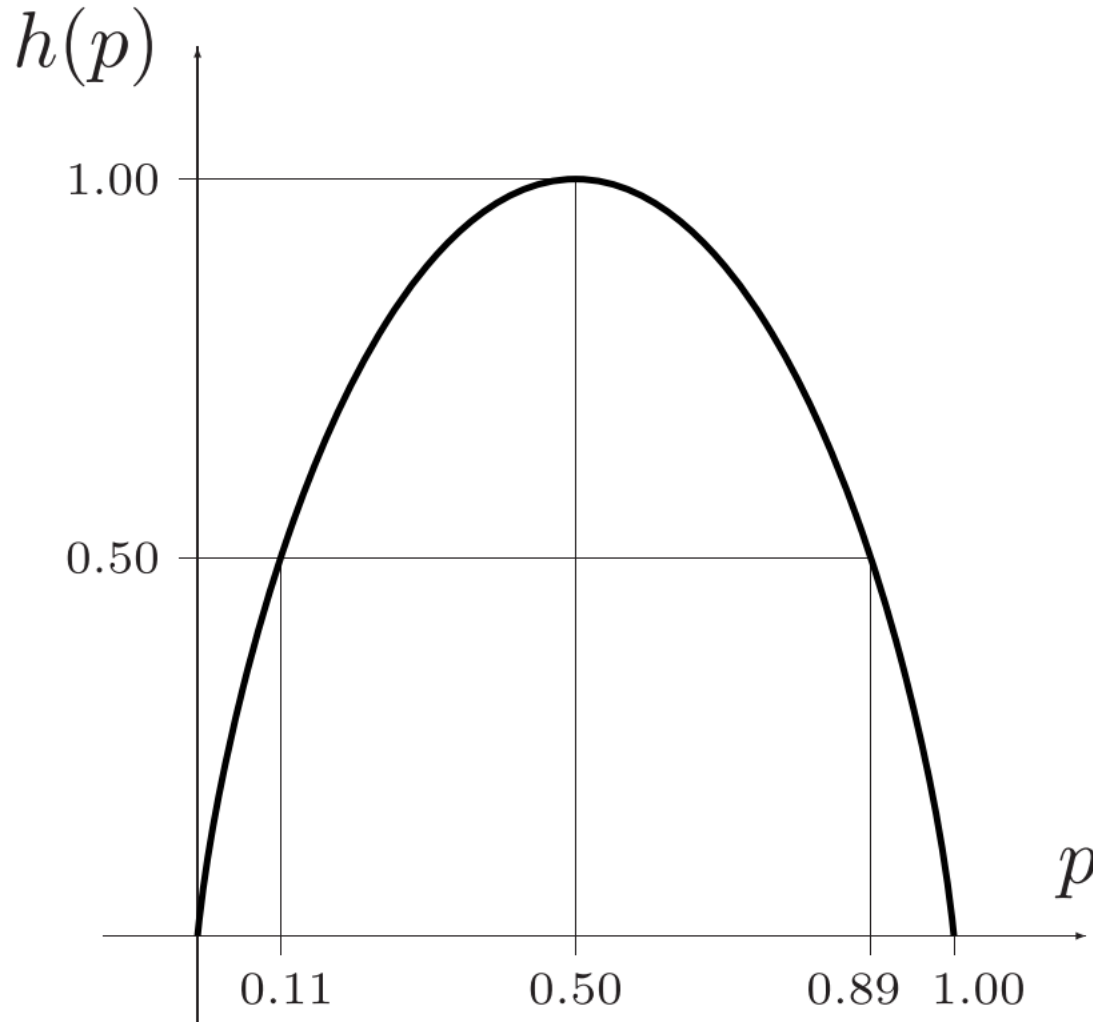
Let X be a binary random variable with outcomes, x_1 and x_2 . Then we have $P_X(x_1) = p$ and $P_X(x_2) = 1 - p$ which yields the uncertainty

$$h(p) \stackrel{\text{def}}{=} -p \log p - (1 - p) \log(1 - p)$$

which we call the *binary entropy function*.



The binary entropy function



Conditional uncertainty

The conditional uncertainty (or conditional entropy) of the discrete random variable X with L outcomes given the discrete random variable Y with M outcomes is the quantity

$$H(X | Y) \stackrel{\text{def}}{=} - \sum_{i=1}^L \sum_{j=1}^M P_{XY}(x_i, y_j) \log P_{X|Y}(x_i | y_j)$$

$P_{XY}(x, y)$ is the joint probability distribution and
 $P_{X|Y}(x | y)$ is the conditional probability distribution.



Conditional uncertainty

For any two discrete random variables X and Y ,

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

with equality if and only if X and Y are independent random variables.

What is then the uncertainty of X if we know Y ?



Bounds for the conditional uncertainty

The conditional uncertainty $H(X | Y)$ of X with L outcomes given Y with M outcomes is lower- and upper-bounded by

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

When does equality hold?



Entropy of joint distributions

Since a pair of random variables is also a random variable it follows that

$$H(XY) = - \sum_{i=1}^L \sum_{j=1}^M P_{XY}(x_i, y_j) \log P_{XY}(x_i, y_j)$$

and

$$\begin{aligned} H(XY) &= H(X) + H(Y | X) \\ &= H(Y) + H(X | Y) \end{aligned}$$



The chain rule for uncertainty

$$H(X_1 X_2 \dots X_N) = H(X_1) + H(X_2 | X_1) \\ + \dots + H(X_N | X_1 X_2 \dots X_{N-1})$$

The uncertainty of the first variable

+the uncertainty of the second given that we know the first

+the uncertainty of the third given that we know the two first

+...



Mutual information

The *information* the random variable Y gives about the random variable X is given by

$$\begin{aligned} I(X;Y) &= H(X) - H(X | Y) \\ &= H(Y) - H(Y | X) \end{aligned}$$

We conclude that the reduction in the uncertainty of one random variable due to the observation of another random variable is symmetric in the two random variables

$$I(X;Y) = I(Y;X)$$



Typical sequences



Typical sequences

All typical long sequences have approximately the same probability and from the law of large numbers it follows that the set of these typical sequences is overwhelmingly probable.

The probability that a long source output sequence is typical is close to one, and, there are approximately

$$2^{nh(p)}$$

typical long sequences.



sequence	probability	
● ● ● ● ●	$1/3 \ 1/3 \ 1/3 \ 1/3 \ 1/3 \Rightarrow$	0.0041
● ● ● ● ○	$1/3 \ 1/3 \ 1/3 \ 1/3 \ 2/3 \Rightarrow$	0.0082
● ● ● ○ ●	$1/3 \ 1/3 \ 1/3 \ 2/3 \ 1/3 \Rightarrow$	0.0082
● ● ● ○ ○	$1/3 \ 1/3 \ 1/3 \ 2/3 \ 2/3 \Rightarrow$	0.0165
● ● ○ ● ●	$1/3 \ 1/3 \ 2/3 \ 1/3 \ 1/3 \Rightarrow$	0.0082
● ● ○ ● ○	$1/3 \ 1/3 \ 2/3 \ 1/3 \ 2/3 \Rightarrow$	0.0165
● ● ○ ○ ●	$1/3 \ 1/3 \ 2/3 \ 2/3 \ 1/3 \Rightarrow$	0.0165
● ● ○ ○ ○	$1/3 \ 1/3 \ 2/3 \ 2/3 \ 2/3 \Rightarrow$	0.0329
● ○ ● ● ●	$1/3 \ 2/3 \ 1/3 \ 1/3 \ 1/3 \Rightarrow$	0.0082
● ○ ● ● ○	$1/3 \ 2/3 \ 1/3 \ 1/3 \ 2/3 \Rightarrow$	0.0165
● ○ ● ○ ●	$1/3 \ 2/3 \ 1/3 \ 2/3 \ 1/3 \Rightarrow$	0.0165
● ○ ● ○ ○	$1/3 \ 2/3 \ 1/3 \ 2/3 \ 2/3 \Rightarrow$	0.0329
● ○ ○ ● ●	$1/3 \ 2/3 \ 2/3 \ 1/3 \ 1/3 \Rightarrow$	0.0165
● ○ ○ ● ○	$1/3 \ 2/3 \ 2/3 \ 1/3 \ 2/3 \Rightarrow$	0.0329
● ○ ○ ○ ●	$1/3 \ 2/3 \ 2/3 \ 2/3 \ 1/3 \Rightarrow$	0.0329
● ○ ○ ○ ○	$1/3 \ 2/3 \ 2/3 \ 2/3 \ 2/3 \Rightarrow$	0.0658
○ ● ● ● ●	$2/3 \ 1/3 \ 1/3 \ 1/3 \ 1/3 \Rightarrow$	0.0082
○ ● ● ● ○	$2/3 \ 1/3 \ 1/3 \ 1/3 \ 2/3 \Rightarrow$	0.0165
○ ● ● ○ ●	$2/3 \ 1/3 \ 1/3 \ 2/3 \ 1/3 \Rightarrow$	0.0165
○ ● ● ○ ○	$2/3 \ 1/3 \ 1/3 \ 2/3 \ 2/3 \Rightarrow$	0.0329
○ ● ○ ● ●	$2/3 \ 1/3 \ 2/3 \ 1/3 \ 1/3 \Rightarrow$	0.0165
○ ● ○ ● ○	$2/3 \ 1/3 \ 2/3 \ 1/3 \ 2/3 \Rightarrow$	0.0329
○ ● ○ ○ ●	$2/3 \ 1/3 \ 2/3 \ 2/3 \ 1/3 \Rightarrow$	0.0329
○ ● ○ ○ ○	$2/3 \ 1/3 \ 2/3 \ 2/3 \ 2/3 \Rightarrow$	0.0658
○ ○ ● ● ●	$2/3 \ 2/3 \ 1/3 \ 1/3 \ 1/3 \Rightarrow$	0.0165
○ ○ ● ● ○	$2/3 \ 2/3 \ 1/3 \ 1/3 \ 2/3 \Rightarrow$	0.0329
○ ○ ● ○ ●	$2/3 \ 2/3 \ 1/3 \ 2/3 \ 1/3 \Rightarrow$	0.0329
○ ○ ● ○ ○	$2/3 \ 2/3 \ 1/3 \ 2/3 \ 2/3 \Rightarrow$	0.0658
○ ○ ○ ● ●	$2/3 \ 2/3 \ 2/3 \ 1/3 \ 1/3 \Rightarrow$	0.0329
○ ○ ○ ● ○	$2/3 \ 2/3 \ 2/3 \ 1/3 \ 2/3 \Rightarrow$	0.0658
○ ○ ○ ○ ●	$2/3 \ 2/3 \ 2/3 \ 2/3 \ 1/3 \Rightarrow$	0.0658
○ ○ ○ ○ ○	$2/3 \ 2/3 \ 2/3 \ 2/3 \ 2/3 \Rightarrow$	0.1317
		0.9998



Properties of typical sequences

Let $\mathcal{T}_\epsilon^{(n)}$ be the set of sequences $\mathbf{x} = x_1x_2 \dots x_n$ such that

$$2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)} \quad (5.43)$$

The set $\mathcal{T}_\epsilon^{(n)}$ is called the typical set and it has the following properties:

1. If $\mathbf{x} \in \mathcal{T}_\epsilon^{(n)}$, then $P_{\mathbf{X}}(\mathbf{x}) \approx 2^{-nH(X)}$.
2. $Pr(\mathcal{T}_\epsilon^{(n)}) > 1 - \epsilon$, for n sufficiently large.
3. $|\mathcal{T}_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.



Longer typical sequences

Let us now choose a smaller ϵ namely $\epsilon = 0.046$ (5 % of $h(1/3)$), and increase the length of the sequences.

Then we obtain the following table:

n	$ \mathcal{T}_\epsilon^{(n)} $	$Pr(\mathcal{T}_\epsilon^{(n)})$
100	$2^{92.6}$	0.660
500	$2^{474.9}$	0.971
1000	$2^{953.4}$	0.998
2000	$2^{1910.3}$	1.000



Typical sequences in text

If we have L letters in our alphabet, then we can compose L^n different sequences that are n letters long.

Only approximately $2^{nH(X)}$, where $H(X)$ is the uncertainty of the language, of these are “meaningful”.

What is meant by “meaningful” is determined by the structure of the language; that is, by its grammar, spelling rules etc.



Typical sequences in text

Only the fraction

$$\frac{2^{nH(X)}}{L^n} = \frac{2^{nH(X)}}{2^{n \log L}} = 2^{-n(\log L - H(X))}$$

which vanishes when n grows provided that $H(X) < \log L$, is "meaningful".

For the English language $H(X)$ is typically 1.5 bits/letter and $\log_2 6 \approx 4.7$ bits/letter.



Structure in text

Shannon illustrated how increasing structure between letters will give better approximations of the English language.

Assuming an alphabet with 27 symbols---26 letters and 1 space---he started with an approximation of the first order.

The symbols are chosen *independently* of each other but with the actual probability distribution (12 % E, 2 % W, etc.):

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA
TH EEI ALHENHTTPA OOBTTVA NAH BRL



Structure in text

Then Shannon continued with the approximation of the second order. The symbols are chosen with the actual *bigram* statistics---when a symbol has been chosen, the next symbol is chosen according to the actual conditional probability distribution:

ON IE ANTSOUTINYS ARE T INCTORE ST BE S
DEAMY ACHIN D ILONASIVE TUCOOWE AT
TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE



Structure in text

The approximation of the third order is based on the *trigram* statistics---when two successive symbols have been chosen, the next symbol is chosen according to the actual conditional probability distribution:

IN NO IST LAT WHEY CRATICT FROURE BIRS
GROCID PONDENOME OF DEMONSTRURES OF THE
REPTAGIN IS REGOACTIONA OF CRE



The principle of source coding

Consider the set of typical long output sequences of n symbols from a source with uncertainty $H(X)$ bits per source symbol.

Since there are fewer than $2^{n(H(X)+\epsilon)}$ typical long sequences in this set, they can be represented by $n(H(X) + \epsilon)$ binary digits; that is, by $H(X) + \epsilon$ binary digits per source symbol.



Summary

- Uncertainty, aka entropy, and the conditional versions of them are calculated from the probabilities of different outcomes of a random variables.
- The binary entropy function is a special case describing the uncertainty of a binary.
- The chain rule is a tool for calculating the entropy of joint distributions.
- Mutual information describes how much the uncertainty (entropy) of a random variable is reduced when some other random variable is observed.
- Typical sequences can be used to estimate, e.g., how efficiently we can source code text or some other sequence of symbols.





LUND
UNIVERSITY