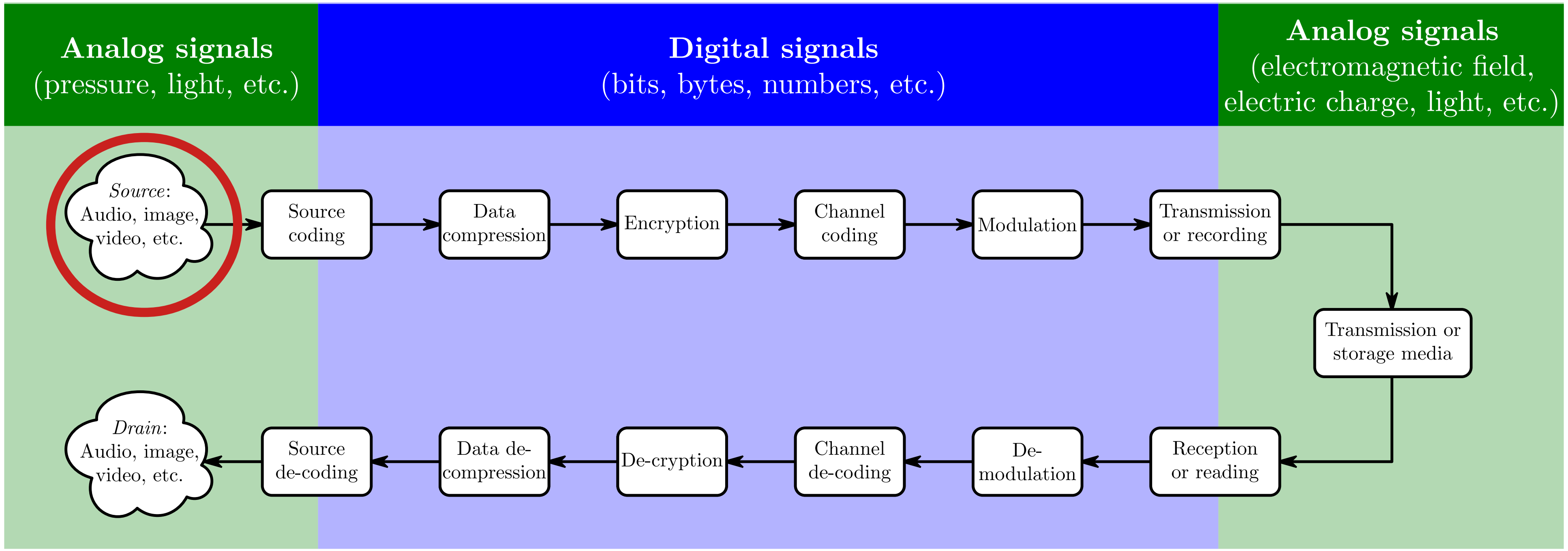
Information Transmission
Chapter 5, Information theroy

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.

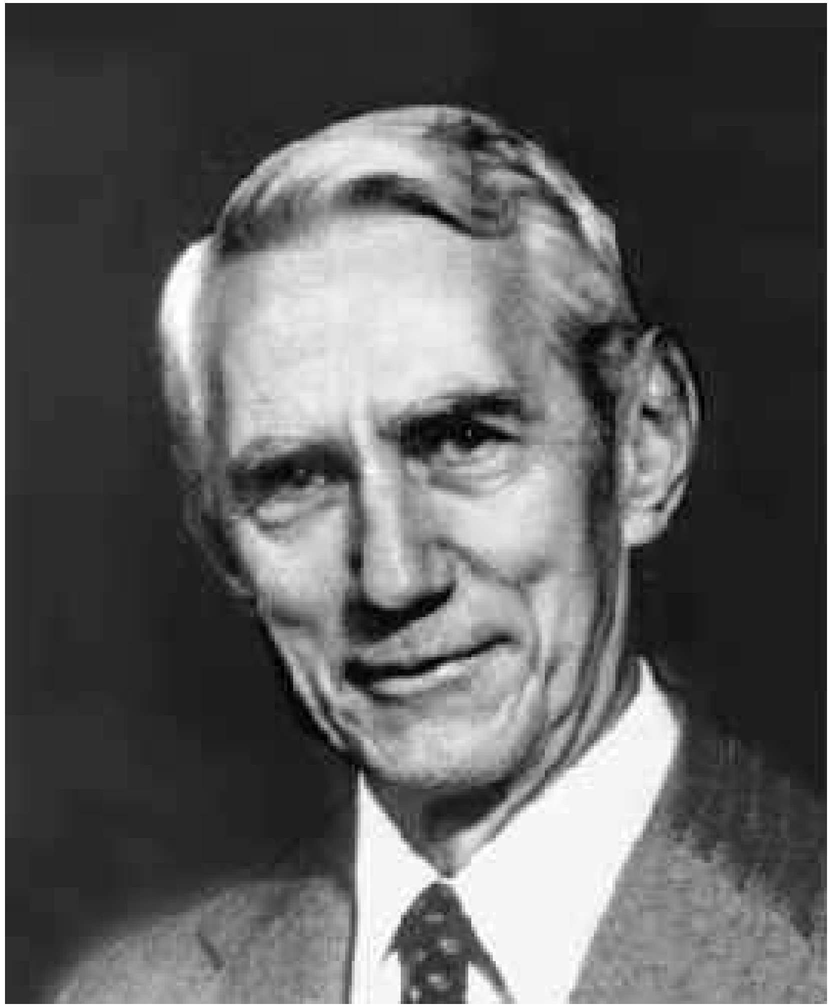
Where are we in the BIG PICTURE?



Properties of information.

Lecture relates to pages 150-166 in textbook.

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 crypthology 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

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

The logarithm is **base 2** when we measure entropy in ***bit***. Sometimes we make this explicit by writing $\mathbf{log_2}$.

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

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

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*.

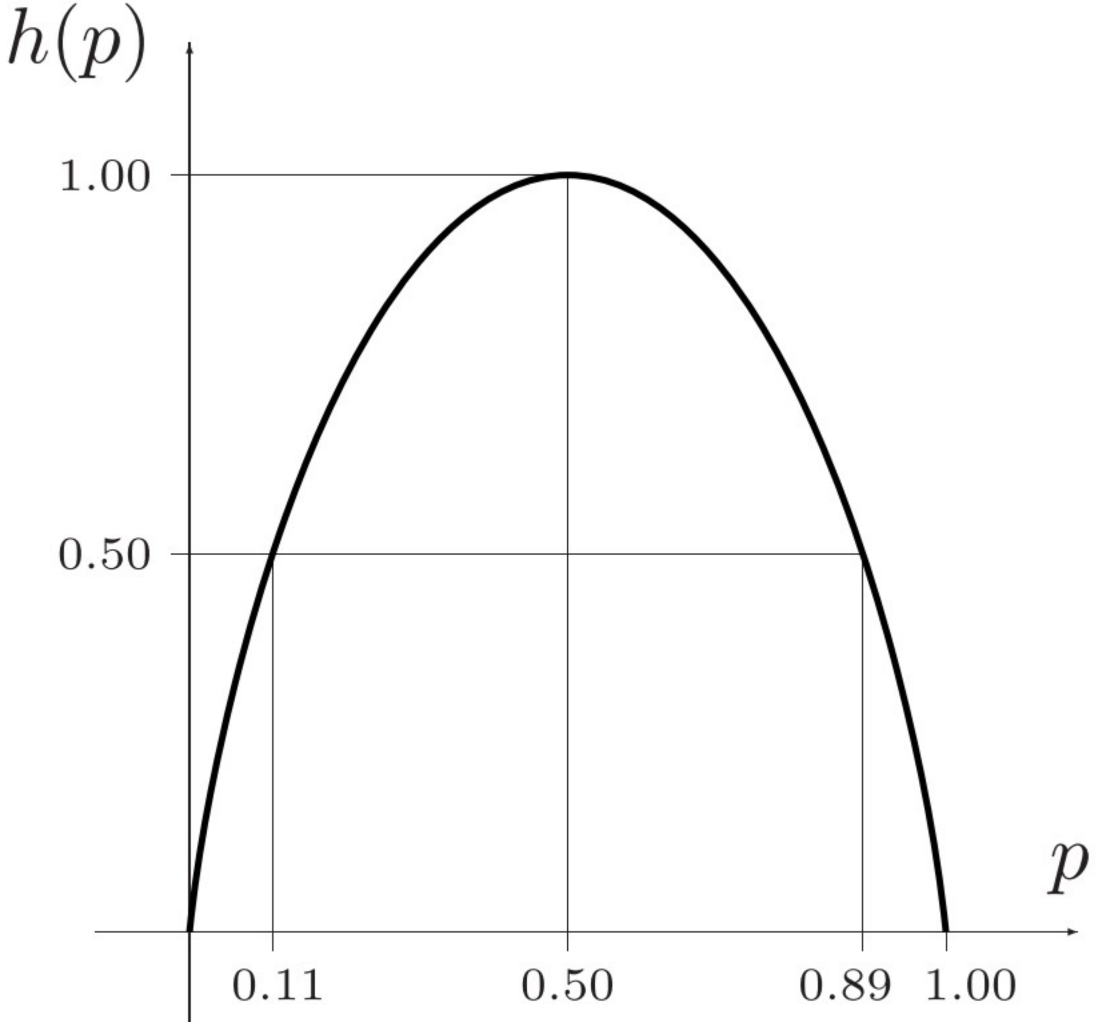
The binary entropy function

Let *X* be a binary random variable with outcomes, x_1 and x_2. When we have $P_X(x_1) = p$ and $P_X(x_2) = 1 - p$, we define the corresponding uncertainty

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

and call *h*(*p*) 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

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

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

Conditional uncertainty

For any two discrete random variables *X* and *Y*,

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

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 $H(X \mid Y)$ of X with L outcomes given Y with M outcomes is lower and upper bounded by

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

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

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

The chain rule for uncertainty

$$
\begin{aligned}
H(X_1X_2\ldots X_N) = H(X_1) + H(X_2 \mid X_1) \\
+ \cdots + H(X_N \mid X_1X_2\cdots X_{N-1})
\end{aligned}
$$

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 first two

+ …

Mutual information

The *information* random variable Y gives

about random variable X is given by

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

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

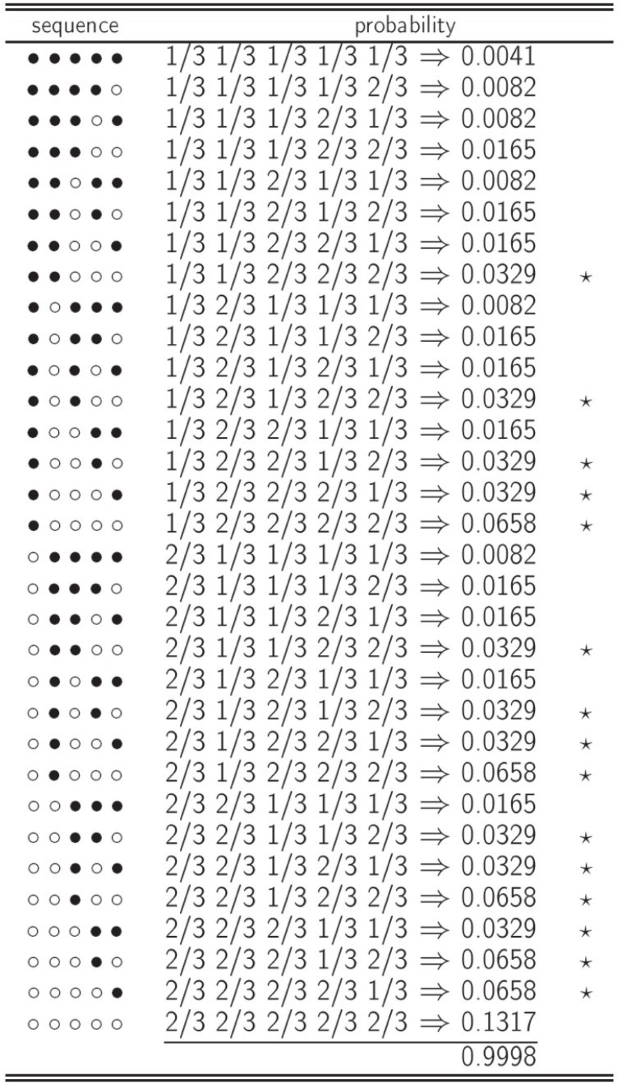
All typical long sequences have approximately the same probability and from the law of large numbers it follows that the set of these typical sequences is overwhelmingly probable.

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

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

typical long sequences.

Example from textbook (draw from urn)



sequence	probability	
● ● ● ● ●	1/3 1/3 1/3 1/3 1/3 ⇒ 0.0041	
● ● ● ● ○	1/3 1/3 1/3 1/3 2/3 ⇒ 0.0082	
● ● ● ○ ●	1/3 1/3 1/3 2/3 1/3 ⇒ 0.0082	
● ● ● ○ ○	1/3 1/3 1/3 2/3 2/3 ⇒ 0.0165	
● ● ○ ● ●	1/3 1/3 2/3 1/3 1/3 ⇒ 0.0082	
● ● ○ ● ○	1/3 1/3 2/3 1/3 2/3 ⇒ 0.0165	
● ● ○ ○ ●	1/3 1/3 2/3 2/3 1/3 ⇒ 0.0165	
● ● ○ ○ ○	1/3 1/3 2/3 2/3 2/3 ⇒ 0.0329	⋆
● ○ ● ● ●	1/3 2/3 1/3 1/3 1/3 ⇒ 0.0082	
● ○ ● ● ○	1/3 2/3 1/3 1/3 2/3 ⇒ 0.0165	
● ○ ● ○ ●	1/3 2/3 1/3 2/3 1/3 ⇒ 0.0165	
● ○ ● ○ ○	1/3 2/3 1/3 2/3 2/3 ⇒ 0.0329	⋆
● ○ ○ ● ●	1/3 2/3 2/3 1/3 1/3 ⇒ 0.0165	
● ○ ○ ● ○	1/3 2/3 2/3 1/3 2/3 ⇒ 0.0329	⋆
● ○ ○ ○ ●	1/3 2/3 2/3 2/3 1/3 ⇒ 0.0329	⋆
● ○ ○ ○ ○	1/3 2/3 2/3 2/3 2/3 ⇒ 0.0658	⋆
○ ● ● ● ●	2/3 1/3 1/3 1/3 1/3 ⇒ 0.0082	
○ ● ● ● ○	2/3 1/3 1/3 1/3 2/3 ⇒ 0.0165	
○ ● ● ○ ●	2/3 1/3 1/3 2/3 1/3 ⇒ 0.0165	
○ ● ● ○ ○	2/3 1/3 1/3 2/3 2/3 ⇒ 0.0329	⋆
○ ● ○ ● ●	2/3 1/3 2/3 1/3 1/3 ⇒ 0.0165	
○ ● ○ ● ○	2/3 1/3 2/3 1/3 2/3 ⇒ 0.0329	⋆
○ ● ○ ○ ●	2/3 1/3 2/3 2/3 1/3 ⇒ 0.0329	⋆
○ ● ○ ○ ○	2/3 1/3 2/3 2/3 2/3 ⇒ 0.0658	⋆
○ ○ ● ● ●	2/3 2/3 1/3 1/3 1/3 ⇒ 0.0165	
○ ○ ● ● ○	2/3 2/3 1/3 1/3 2/3 ⇒ 0.0329	⋆
○ ○ ● ○ ●	2/3 2/3 1/3 2/3 1/3 ⇒ 0.0329	⋆
○ ○ ● ○ ○	2/3 2/3 1/3 2/3 2/3 ⇒ 0.0658	⋆
○ ○ ○ ● ●	2/3 2/3 2/3 1/3 1/3 ⇒ 0.0329	⋆
○ ○ ○ ● ○	2/3 2/3 2/3 1/3 2/3 ⇒ 0.0658	⋆
○ ○ ○ ○ ●	2/3 2/3 2/3 2/3 1/3 ⇒ 0.0658	⋆
○ ○ ○ ○ ○	2/3 2/3 2/3 2/3 2/3 ⇒ 0.1317	
	0.9998	

● - probability 1/3
○ - probability 2/3

Number of typical sequences should be about:

$$2^{nh(p)} = 2^{5h(1/3)} = 2^{5 \cdot 0.9183} \approx 24$$

Sequences with “observed uncertainty” within 15% of h(1/3) (probability between 0.027 and 0.068):

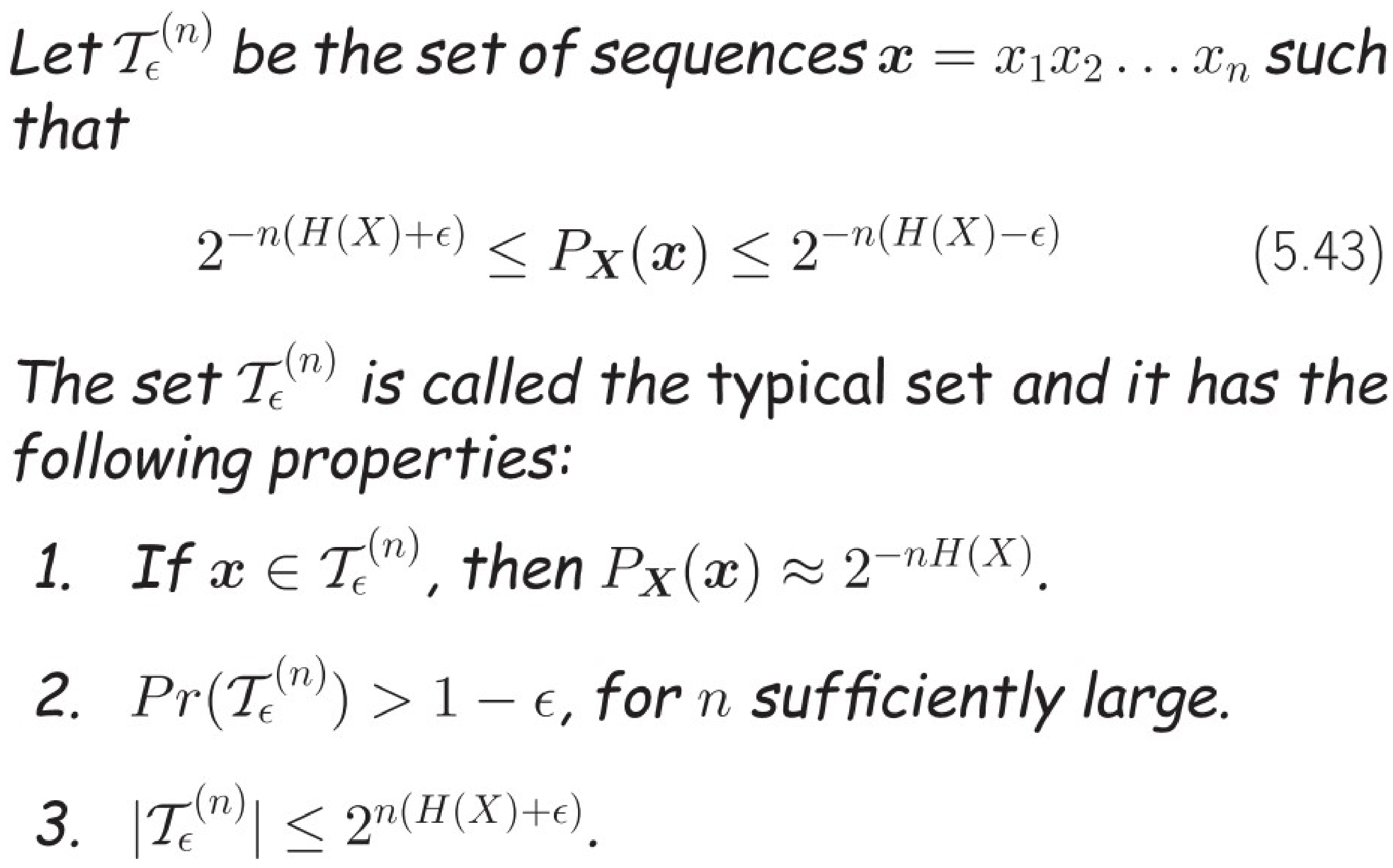
15 (the ones marked with stars)

Why the large discrepancy?

Only valid for “long” sequences.

… but the 15 sequences are less than 1/2 of all sequences and contain about 2/3 of all probability.

Properties of typical sequences



Let $\mathcal{T}_\epsilon^{(n)}$ be the set of sequences $\boldsymbol{x} = x_1 x_2 \ldots x_n$ such that

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

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

1. If $\boldsymbol{x} \in \mathcal{T}_\epsilon^{(n)}$, then $P_{\boldsymbol{X}}(\boldsymbol{x}) \approx 2^{-nH(X)}$.
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	$\lvert\mathcal{T}_\epsilon^{(n)}\rvert$	$Pr(\mathcal{T}_\epsilon^{(n)})$
100	$2^{92.6}$	0.660
500	$2^{474.9}$	0.971
1000	$2^{953.4}$	0.998
2000	$2^{1910.3}$	1.000

Note: In the first example with length-five sequences we had a wider tolerance of 15% of h(1/3), and captured 2/3 of the probability in our typical sequences.

With this tighter tolerance we need sequences of length 100 to capture 2/3 of the total probability in the typical sequences.

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 a fraction

$$\frac{2^{nH(X)}}{L^n} = \frac{2^{nH(X)}}{2^{n\log_2 L}} = 2^{-n(\log_2 L - H(X))},$$

which vanishes when *n* grows provided that $H(X) < \log_2 L$, is ”meaningful” text of length *n* letters.

For the English language *H*(*X*) is typically 1.5 bits/letter and $\log_2 L = \log_2 26 \approx 4.7$ 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 one 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 specal 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.

* SIGILLVM · VNIVERSITATIS · GOTHORVM · CAROLINÆ

AD VTRVMQVE

1666

LUND
UNIVERSITY