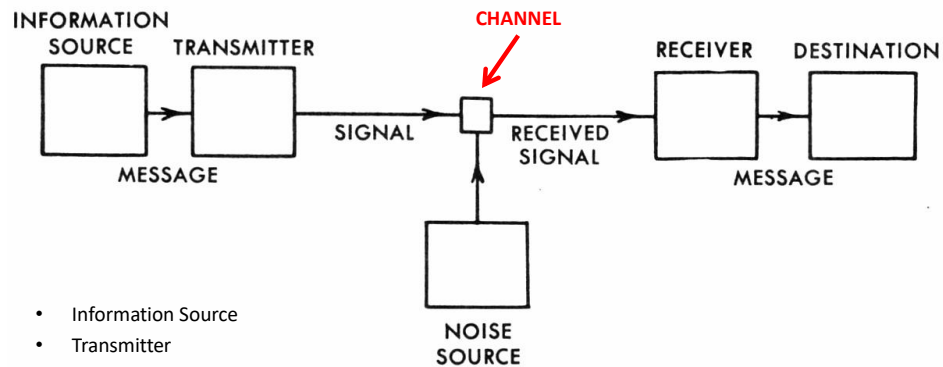# Introduction to Information Theory

Part 2

1

A General Communication System

- Information Source
- Transmitter
- Channel
- Receiver
- Destination

2

1

Definition of Information

- ➢ Information is quantified using probabilities.
- ➢ Given a finite set of possible messages, associate a probability with each message.
- ➢ A message with low probability represents more information than one with high probability.

Definition of Information:

$$I = \log_2\left(\frac{1}{p}\right) = -\log_2(p)$$

Where p is the probability of the message
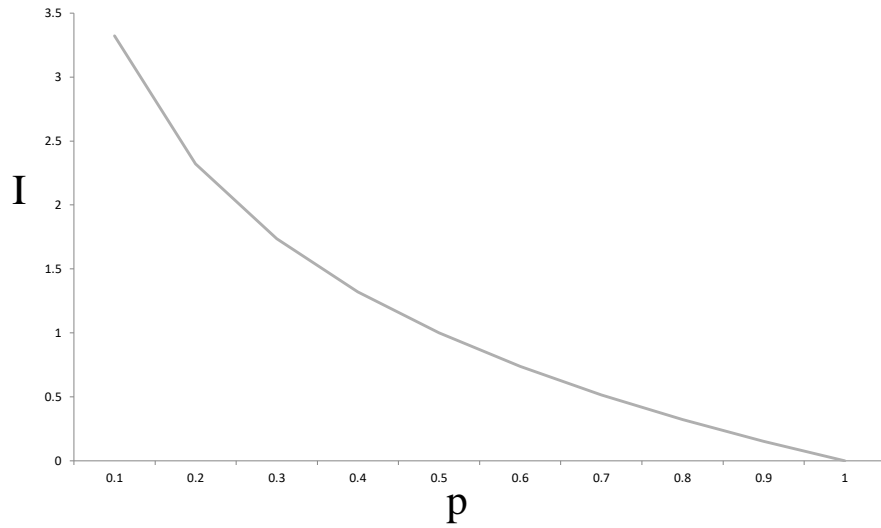Base 2 is used for the logarithm so I is measured in **bits**

3

Example: Information in a coin flip

$$p(\text{HEADS}) = \frac{1}{2}$$
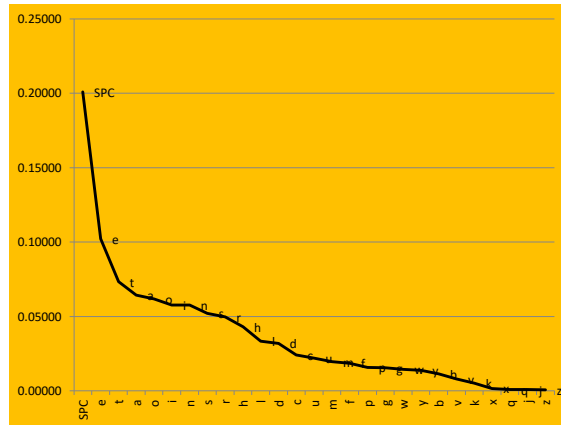
$$I = -\log_2\left(\frac{1}{2}\right) = 1 \text{ bit}$$

4

Information Content

Example: Text Analysis

a	0.06428
b	0.01147
c	0.02413
d	0.03188
e	0.10210
f	0.01842
g	0.01543
h	0.04313
i	0.05767
j	0.00082
k	0.00514
l	0.03338
m	0.01959
n	0.05761
o	0.06179
p	0.01571
q	0.00084
r	0.04973
s	0.05199
t	0.07327
u	0.02201
v	0.00800
w	0.01439
x	0.00162
y	0.01387
z	0.00077
SPC	0.20096

Example Text Analysis

Letter	Freq.	I
a	0.06428	3.95951
b	0.01147	6.44597
c	0.02413	5.37297
d	0.03188	4.97116
e	0.10210	3.29188
f	0.01842	5.76293
g	0.01543	6.01840
h	0.04313	4.53514
i	0.05767	4.11611
j	0.00082	10.24909
k	0.00514	7.60474
l	0.03338	4.90474
m	0.01959	5.67385
n	0.05761	4.11743
o	0.06179	4.01654
p	0.01571	5.99226
q	0.00084	10.21486
r	0.04973	4.32981
s	0.05199	4.26552
t	0.07327	3.77056
u	0.02201	5.50592
v	0.00800	6.96640
w	0.01439	6.11899
x	0.00162	9.26697
y	0.01387	6.17152
z	0.00077	10.34877
SPC	0.20096	2.31502

7

Informattion, $I = \log(1/p)$

Some properties of I

1. $I(p) \geq 0$
   Information is non-negative.

2. $I(p_1 * p_2) = I(p_1) + I(p_1)$
   Information we get from observing two independent events occurring is the sum of two information(s).

3. $I(p)$ is monotonic and continuous in p
   Slight changes in probability incur slight changes in information.

4. $I(p = 1) = 0$
   We get zero information from an event whose probability is 1.

8

4

Entropy

➢ Information (I) is associated with known events/messages that have occurred.

➢ Entropy is a measure of information we expect to receive in the future.

➢ It is the average information w.r.to all possible outcomes

9

Entropy

➢ Information (I) is associated with known events/messages that have occurred.

➢ Entropy is a measure of information we expect to receive in the future.

➢ It is the average information w.r.to all possible outcomes

$$p_1 I_1 + p_2 I_2 + p_3 I_3 + \cdots$$

10

Definition of Entropy

➢ Information (I) is associated with known events/messages
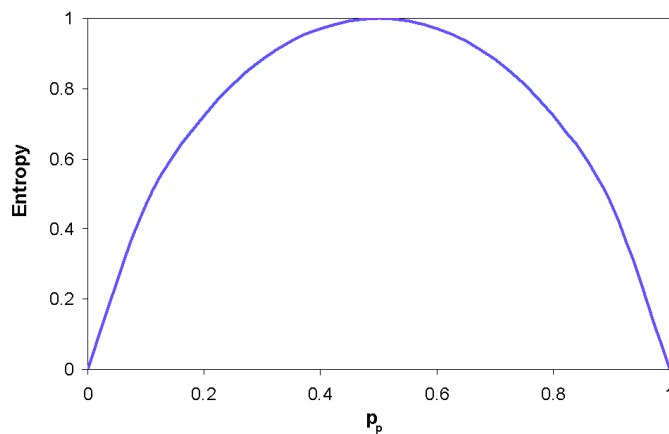➢ Entropy (H) is the average information w.r.to all possible outcomes

$$H(X) = \sum_x p_x \log \frac{1}{p_x}$$

➢ H is also measured in bits

11

Entropy (2 outcomes: $p, 1 - p$)

$$H(p) = p \log\left(\frac{1}{p}\right) + (1 - p) \log(\frac{1}{1-p})$$

12

Examples: Weather

- 2 event source: (Sunny, Cloudy)

$$p_{sunny} = \frac{7}{8}$$
$$p_{cloudy} = \frac{1}{8}$$

- 3 – event source (Sunny, Cloudy, Precip)

  Bryn Mawr: (207, 119, 39)

  Let us compute this.

13

Example Text Analysis

Letter	Freq.	I, h(pi)
a	0.06428	3.95951
b	0.01147	6.44597
c	0.02413	5.37297
d	0.03188	4.97116
e	0.10210	3.29188
f	0.01842	5.76293
g	0.01543	6.01840
h	0.04313	4.53514
i	0.05767	4.11611
j	0.00082	10.24909
k	0.00514	7.60474
l	0.03338	4.90474
m	0.01959	5.67385
n	0.05761	4.11743
o	0.06179	4.01654
p	0.01571	5.99226
q	0.00084	10.21486
r	0.04973	4.32981
s	0.05199	4.26552
t	0.07327	3.77056
u	0.02201	5.50592
v	0.00800	6.96640
w	0.01439	6.11899
x	0.00162	9.26697
y	0.01387	6.17152
z	0.00077	10.34877
SPC	0.20096	2.31502

$$H(X) = \sum_{x} p_x \log \frac{1}{p_x} = 4.047$$

14

7

Entropy: Properties
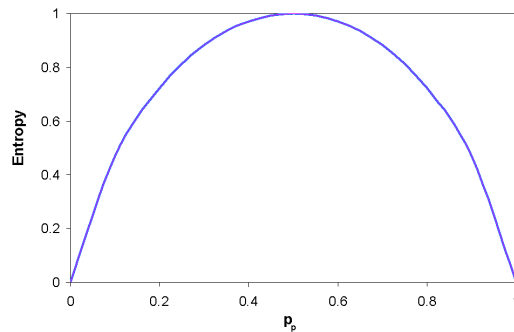
$$H(X) \geq 0 \qquad H(X_n) \leq \log_2(n) \qquad H(S, T) = H(S) + H(T)$$

Entropy is maximized if p is uniform.   Additive Property

15

Entropy of S^n

- S is a source with k independent events and H(S) = e
  e.g. S = [H, T]
  H, H, T, H, T, H, …
  H(S) = 1

- S^2 is a source consisting of two observations of events from S
  e.g. S = [H, T]
  TH, TT, HH, HH, TT, HT, …

  then, $H(S^2) = 2\,H(S)$

- In general,
  $$H(S^n) = n\,H(S)$$

16

Entropy of things…

- Entropy of English text is approx 1.5 bits/letter

- Entropy of the human genome <= 2 bits

- Entropy of a black hole is ¼ of the area of the outer event horizon.

- Value of information in economics is defined in terms of entropy. E.g. Scarcity

$$V(X) = \sum_{i=1}^{n} p_i(-\log_b(p_i))$$

17

bit versus bit - **Two meanings**

- bit as measure of information/entropy
- bit as a binary digit

    e.g.   01001101   is six bits long
    weather   01001101   8 days sunny/cloudy(0/1)
                                                information is less than 8 bits

    Information represented as decimal digits
    log(10) = 3.32, thus the string 32767 has 6 * 3.32 = 19.92 bits of information

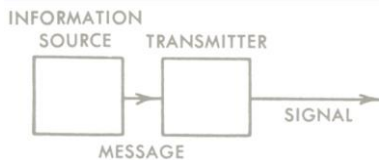   26 letter-alphabet has average information, log(26) = 4.7 bits

- Bits needed to store n symbols matches entropy in bits only when all symbols are equally likely and are mutually independent.

18

So, what is Entropy good for???

- Provides the foundation for techniques for
  - Compression
  - Searching in data
  - Encryption
  - Correcting communication errors
  - Extracting information from data
  - Economic value of information
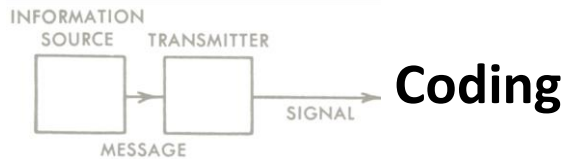 - Biological information
  - Quantum information
  - Etc.

19

Coding

INFORMATION
SOURCE   TRANSMITTER

SIGNAL

MESSAGE

- An information source, S has m events
- Thus, m symbols are to be transmitted: $s_1, s_2, s_3, \ldots, s_m$

20

Coding

- An information source, S has m events
- Thus, m symbols are to be transmitted: $s_1, s_2, s_3, \ldots, s_m$

- A **code** is an assignment of **codewords** to source symbols
- Codewords are made up of characters from a **code alphabet**

e.g. $S = \{SUNNY, PRECIP, RAINY\}$

code alphabet = {0, 1}

Code:

SUNNY	$\rightarrow$	0
PRECIP	$\rightarrow$	01
RAINY	$\rightarrow$	010

21

Coding: Basics

- Events of an information source: $s_1, s_2, \ldots, s_n$

- A **code** is made up of **codewords** from a **code alphabet**
(e.g. {0, 1}, {., -}, etc.)

22

Coding: Basics

- **Block code**: When all codes have the same length. For example, ASCII, Unicode, etc.

- **Average Word Length**: $L = \sum_{i=1}^{m} p_i l_i$

- **Singular** (not unique) codes
- **Nonsingular** (unique) codes
- **instantaneous** codes

23

Coding: Basics

- **Block code**: When all codes have the same length. For example, ASCII ($l = 8$).

- **Average Word Length**: $L = \sum_{i=1}^{m} p_i l_i$

Code length is important!

Useless code! • **Singular** (not unique) codes

Short codewords preferred to long ones.

- **Nonsingular** (unique) codes
- **instantaneous** codes

24

Example Code

Source Symbol	Singular Code	Nonsingular Code
A	00	0
B	10	10
C	01	00
D	10	01

25

Example Code

Source Symbol	Singular Code	Nonsingular Code
A	00	0
B	10	10
C	01	00
D	10	01

In practice, nonsingularity is not sufficient.

e.g.   receiver gets: 0010

ADA?
CD?
AAB?

26

Nonsingular, Instantaneous, Block Code

Source Symbol	Nonsingular Code
A	00
B	01
C	10
D	10

e.g.   receiver gets: 01101100

27

Comma Codes & Capital Codes

Source Symbol	Comma Code	Capital Code
A	0	0
B	10	01
C	110	011
D	1110	0111

One of these is instantaneous.

e.g.   receiver gets: 01011100

receiver gets: 00101110

28

14

Example

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

29

Example

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

$$H = 0.3 \log\left(\frac{1}{0.3}\right) + 0.2 \log\left(\frac{1}{0.2}\right) * 3 + 0.1 \log\left(\frac{1}{0.1}\right) = 2.246$$

30

Example

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

?

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

$$H = 0.3 \log\left(\frac{1}{0.3}\right) + 0.2 \log\left(\frac{1}{0.2}\right) * 3 + 0.1 \log\left(\frac{1}{0.1}\right) = 2.246$$

Is there a relationship between L and H?

31

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L \leq H(S^n) + 1$$

32

Average Code Length & Entropy

- Average length bounds: $\quad H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

$$nH(S) \leq L < nH(S) + 1$$

33

Average Code Length & Entropy

- Average length bounds: $\quad H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L \leq H(S^n) + 1$$

$$nH(S) \leq L \leq nH(S) + 1$$

$$H(S) \leq \frac{L}{n} \leq H(S) + \frac{1}{n}$$

This is for instantaneous binary codes.

34

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L \leq H(S^n) + 1$$

$$nH(S) \leq L \leq nH(S) + 1$$

$$H(S) \leq \frac{L}{n} \leq H(S) + \frac{1}{n}$$

$$\lim_{n \to \infty} \frac{L_n}{n} = H$$

H is the entropy of source S
n is the length of symbol sequences
L_n is the avg. length of codewords

35

Shannon's First Theorem

- By coding sequences of independent symbols (in S^n), it is possible to construct codes such that

$$\lim_{n \to \infty} \frac{L_n}{n} = H$$

The price paid for such improvement is increased coding complexity (due to increased n) and increased delay in coding.

36

Question

- Is there a **coding algorithm** that produces codes such that it achieves Shannon limit?

    L = H?

Yes!

Huffman's algorithm (**Huffman Coding**) produces a code with average length L as close as possible to source code entropy, H.

37

References

- Eugene Chiu, Jocelyn Lin, Brok Mcferron, Noshirwan Petigara, Satwiksai Seshasai: *Mathematical Theory of Claude Shannon: A study of the style and context of his work up to the genesis of information theory. MIT* 6.933J / STS.420J The Structure of Engineering Revolutions
- Luciano Floridi, 2010: *Information: A Very Short Introduction*, Oxford University Press, 2011.
- Luciano Floridi, 2011: *The Philosophy of Information*, Oxford University Press, 2011.
- James Gleick, 2011: *The Information: A History, A Theory, A Flood*, Pantheon Books, 2011.
- Zhandong Liu , Santosh S Venkatesh and Carlo C Maley, 2008: *Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples*, *BMC Genomics* 2008, **9:**509
- David Luenberger, 2006: *Information Science*, Princeton University Press, 2006.
- David J.C. MacKay, 2003: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
- Claude Shannon & Warren Weaver, 1949: *The Mathematical Theory of Communication*, University of Illinois Press, 1949.
- W. N. Francis and H. Kucera: *Brown University Standard Corpus of Present-Day American English*, Brown University, 1967.

38