Introduction to Information Theory

Part 2
A General Communication System

- Information Source
- Transmitter
- Channel
- Receiver
- Destination
Information: Definition

- 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:

\[ I(p) = \log \left( \frac{1}{p} \right) = -\log(p) \]

Where \( p \) is the probability of the message
Base 2 is used for the logarithm so \( I \) is measured in \( \text{bits} \)
Trits for base 3, \( \text{nats} \) for base \( e \), \( \text{Hartleys} \) for base 10...
\[ I(p) = \log\left(\frac{1}{p}\right) = -\log(p) \]

Some properties of \( I \)

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

2. \( I(p_1 \times p_2) = I(p_1) + I(p_1) \)
   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.
Example: Information in a coin flip

$p_{HEADS} = 1/2$

$I_{HEADS} = -\log(1/2) = 1\text{bit}$
Independent Events: 2 Coin flips

- There are four possibilities: HH, HT, TH, TT

\[ I_{HH} = \log \left( \frac{1}{p_H \times p_H} \right) = \log \left( \frac{1}{1/4} \right) = \log(4) = 2 \]

i.e. Additive property:

\[ I_{AB} = - \log(p_A p_B) = - \log(p_A) - \log(p_B) \]

\[ I_{AB} = I_A + I_B \]
Example: Text Analysis

<table>
<thead>
<tr>
<th>Letter</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.06428</td>
</tr>
<tr>
<td>b</td>
<td>0.01147</td>
</tr>
<tr>
<td>c</td>
<td>0.02413</td>
</tr>
<tr>
<td>d</td>
<td>0.03188</td>
</tr>
<tr>
<td>e</td>
<td>0.10210</td>
</tr>
<tr>
<td>f</td>
<td>0.01842</td>
</tr>
<tr>
<td>g</td>
<td>0.01543</td>
</tr>
<tr>
<td>h</td>
<td>0.04313</td>
</tr>
<tr>
<td>i</td>
<td>0.05767</td>
</tr>
<tr>
<td>j</td>
<td>0.00082</td>
</tr>
<tr>
<td>k</td>
<td>0.00514</td>
</tr>
<tr>
<td>l</td>
<td>0.03338</td>
</tr>
<tr>
<td>m</td>
<td>0.01959</td>
</tr>
<tr>
<td>n</td>
<td>0.05761</td>
</tr>
<tr>
<td>o</td>
<td>0.06179</td>
</tr>
<tr>
<td>p</td>
<td>0.01571</td>
</tr>
<tr>
<td>q</td>
<td>0.00084</td>
</tr>
<tr>
<td>r</td>
<td>0.04973</td>
</tr>
<tr>
<td>s</td>
<td>0.05199</td>
</tr>
<tr>
<td>t</td>
<td>0.07327</td>
</tr>
<tr>
<td>u</td>
<td>0.02201</td>
</tr>
<tr>
<td>v</td>
<td>0.00800</td>
</tr>
<tr>
<td>w</td>
<td>0.01439</td>
</tr>
<tr>
<td>x</td>
<td>0.00162</td>
</tr>
<tr>
<td>y</td>
<td>0.01387</td>
</tr>
<tr>
<td>z</td>
<td>0.00077</td>
</tr>
<tr>
<td>SPC</td>
<td>0.20096</td>
</tr>
</tbody>
</table>
## Example: Text Analysis

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

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

Given, $P = \{p_1, p_2, \ldots, p_3\}$

$$H(P) = \sum_i p_i \log\left(\frac{1}{p_i}\right)$$

Characterizes an information source.
Example: A 3-event Source

\[ A = \{a_1, a_2, a_3\} \]

\[ P = \{p, p_2, p_3\} = \left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{4} \right\} \]

\[ H(P) = \frac{1}{2} \log(2) + \frac{1}{4} \log(4) + \frac{1}{4} \log(4) \]

\[ = \frac{1}{2} + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = \frac{3}{2} = 1.5 \text{ bits} \]
Example: Text Analysis

<table>
<thead>
<tr>
<th>Letter</th>
<th>Freq.</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.06428</td>
<td>3.95951</td>
</tr>
<tr>
<td>b</td>
<td>0.01147</td>
<td>6.44597</td>
</tr>
<tr>
<td>c</td>
<td>0.02413</td>
<td>5.37297</td>
</tr>
<tr>
<td>d</td>
<td>0.03188</td>
<td>4.97116</td>
</tr>
<tr>
<td>e</td>
<td>0.10210</td>
<td>3.29188</td>
</tr>
<tr>
<td>f</td>
<td>0.01842</td>
<td>5.76293</td>
</tr>
<tr>
<td>g</td>
<td>0.01543</td>
<td>6.01840</td>
</tr>
<tr>
<td>h</td>
<td>0.04313</td>
<td>4.53514</td>
</tr>
<tr>
<td>i</td>
<td>0.05767</td>
<td>4.11611</td>
</tr>
<tr>
<td>j</td>
<td>0.00082</td>
<td>10.24909</td>
</tr>
<tr>
<td>k</td>
<td>0.00514</td>
<td>7.60474</td>
</tr>
<tr>
<td>l</td>
<td>0.03338</td>
<td>4.90474</td>
</tr>
<tr>
<td>m</td>
<td>0.01959</td>
<td>5.67385</td>
</tr>
<tr>
<td>n</td>
<td>0.05761</td>
<td>4.11743</td>
</tr>
<tr>
<td>o</td>
<td>0.06179</td>
<td>4.01654</td>
</tr>
<tr>
<td>p</td>
<td>0.01571</td>
<td>5.99226</td>
</tr>
<tr>
<td>q</td>
<td>0.00084</td>
<td>10.21486</td>
</tr>
<tr>
<td>r</td>
<td>0.04973</td>
<td>4.32981</td>
</tr>
<tr>
<td>s</td>
<td>0.03195</td>
<td>4.26552</td>
</tr>
<tr>
<td>t</td>
<td>0.07327</td>
<td>3.77056</td>
</tr>
<tr>
<td>u</td>
<td>0.02201</td>
<td>5.50592</td>
</tr>
<tr>
<td>v</td>
<td>0.00800</td>
<td>6.96640</td>
</tr>
<tr>
<td>w</td>
<td>0.01439</td>
<td>6.11899</td>
</tr>
<tr>
<td>x</td>
<td>0.00162</td>
<td>9.26697</td>
</tr>
<tr>
<td>y</td>
<td>0.01387</td>
<td>6.17152</td>
</tr>
<tr>
<td>z</td>
<td>0.00077</td>
<td>10.34877</td>
</tr>
<tr>
<td>SPC</td>
<td>0.20096</td>
<td>2.31502</td>
</tr>
</tbody>
</table>

\[
H(P) = \sum_i p_i \log \left(\frac{1}{p_i}\right) = 4.047
\]

Aka, First-Order Entropy.
Entropy (2 outcomes)
Entropy: Properties

1. $H(P) \geq 0$

2. $H(P) \leq \log(n)$
   Entropy is maximized if $P$ is uniform.

3. $H(S, T) = H(S) + H(T)$
   Additive property for independent events.

4. $H(S, T) \leq H(S) + H(T)$
   If $S$ and $T$ are not independent.
Entropy of things...

- Entropy of English text is approx 1.5 bits
- Entropy of the human genome $\leq 2$ bits
- Entropy of a black hole is $\frac{1}{4}$ of the area of the outer event horizon.
- Value of information in economics is defined in terms of entropy. E.g. Scarcity

$$V(X) = \sum_{i=1}^{n} p_i(-\log_b(p_i))$$
Entropy: What about it?

• Does $H(P)$ have a maximum? Where?

• Is entropy a good name for this stuff? How is it related to entropy in thermodynamics?

• How does entropy help in communication? What else can we do with it?

• Why use the letter $H$? 😊
Entropy is closely related to the design of efficient codes for random sources.

Provides foundations for techniques of compression, data search, encryption, correction of communication errors, etc.

Essential to the study of life sciences, economics, etc.
Coding: Basics

• Events of an information source: $s_1, s_2, ..., s_m$

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

• A **code** is an assignment of codewords to source symbols.
Coding: Basics

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

• **Average Word Length**:

\[ L = \sum_{i=1}^{m} p_i l_i \]

More generally,

\[ L_n = \frac{1}{n} \sum_{i=1}^{m} p_i l_i \]

• A code is **efficient** if it has the smallest average word length. (Turns out entropy is the benchmark...)
Coding: Basics

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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Singular Code</th>
<th>Nonsingular Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>00</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>01</td>
<td>00</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>01</td>
</tr>
</tbody>
</table>
Coding: Basics

- **Singular** (not unique) codes
- **Nonsingular** (unique) codes
- **instantaneous** codes
  (every word can be decoded as soon as it is received)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Singular Code</th>
<th>Nonsingular Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>00</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>01</td>
<td>00</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>01</td>
</tr>
</tbody>
</table>
Example: Avg. Code Length (L)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>p</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>00</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>11</td>
</tr>
<tr>
<td>D</td>
<td>0.2</td>
<td>010</td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>011</td>
</tr>
</tbody>
</table>

$L = (0.3 \times 2) + (0.2 \times 2) + (0.2 \times 2) + (0.2 \times 3) + (0.1 \times 3) = 2.3$
Example: Source Entropy (H)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>p</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>00</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>11</td>
</tr>
<tr>
<td>D</td>
<td>0.2</td>
<td>010</td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>011</td>
</tr>
</tbody>
</table>

\[
L = (0.3 \times 2) + (0.2 \times 2) + (0.2 \times 2) + (0.2 \times 3) + (0.1 \times 3) = 2.3
\]

\[
H = 0.3 \log \left( \frac{1}{0.3} \right) + 0.2 \log \left( \frac{1}{0.2} \right) \times 3 + 0.1 \log \left( \frac{1}{0.1} \right) = 2.246
\]
Example: $L \& H$

<table>
<thead>
<tr>
<th>Symbol</th>
<th>p</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>00</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>11</td>
</tr>
<tr>
<td>D</td>
<td>0.2</td>
<td>010</td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>011</td>
</tr>
</tbody>
</table>

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

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

Is there a relationship between $L$ and $H$?
Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$

- Grouping $n$ symbols together:

  $$H(S^n) \leq L < H(S^n) + 1$$
Average Code Length & Entropy

• Average length bounds: $H \leq L < H + 1$

• Grouping $n$ symbols together:

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

$nH(S) \leq L < nH(S) + 1$
Average Code Length & Entropy

• Average length bounds: $H \leq L < H + 1$

• Grouping $n$ symbols together:

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

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

$$H(S) \leq \frac{L}{n} < H(S) + \frac{1}{n}$$
Average Code Length & Entropy

• Average length bounds: $H \leq L < H + 1$

• Grouping $n$ symbols together:

\[ H(S^n) \leq L < H(S^n) + 1 \]
\[ nH(S) \leq L < nH(S) + 1 \]
\[ H(S) \leq \frac{L}{n} < H(S) + \frac{1}{n} \]
Average Code Length & Entropy

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

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

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

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

$$\lim_{n \to \infty} \frac{L_n}{n} = H$$
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.
Entropy & Coding: Central Ideas

• Use short codes for highly likely events. This shortens the average length of coded messages.

• Code several events at a time. Provides greater flexibility in code design.
Data Compression: Huffman Coding

A  0.3
B  0.2
C  0.2
D  0.2
E  0.1
Huffman Coding: Reduction Phase

A  0.3  0.3  0.4  0.6
B  0.2  0.3  0.3  0.4
C  0.2  0.2  0.3
D  0.2  0.2
E  0.1
Huffman Coding: Splitting Phase

A  0.3  00  0.3  00  0.4  1  0.6  0
B  0.2  10  0.3  01  0.3  00  0.4  1
C  0.2  11  0.2  10  0.3  01
D  0.2  010  0.2  11
E  0.1  011
### Huffman Coding: Splitting Phase

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>00</td>
<td>0.3</td>
<td>00</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>10</td>
<td>0.3</td>
<td>01</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>11</td>
<td>0.2</td>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td>D</td>
<td>0.2</td>
<td>010</td>
<td>0.2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>011</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ H = \sum p_i \log_2 \frac{1}{p_i} = 2.246 \]

\[ L = (0.3 \times 2) + (0.2 \times 2) + (0.2 \times 2) + (0.2 \times 3) + (0.1 \times 3) = 2.3 \]
Huffman Codes

• Nonsingular
• Instantaneous
• Efficient
• Non-unique
• Powers of a source lead closer to $H$
• Requires knowledge of symbol probabilities
Design Huffman Codes

- \( S = \{A, B\}, P = \{0.75, 0.25\} \)

- \( S = \{AA, AB, BA, BB\} \)

- \( S = \{AAA, AAB, ABA, BAA, ABB, BAB, BBA, BBB\} \)
References


• 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


\[ H(S, T) = H(S) + H(T) \]

Additive property.

S & T are independent sources,

\[ H(S, T) = - \sum_{s \in S, t \in T} p_s p_t \log(p_s p_t) \]

\[ = - \sum_{s \in S, t \in T} p_s p_t [\log(p_s) + \log(p_t)] \]

\[ = - \sum_{t \in T} p_t \left[ \sum_{s \in S} p_s \log(p_s) \right] - \sum_{s \in S} p_s \left[ \sum_{t \in T} p_t \log(p_t) \right] \]

\[ = H(S) + H(T) \]