Information Theory

Mohamed Hamada

Software Engineering Lab
The University of Aizu

Email: hamada@u-aizu.ac.jp
URL: http://www.u-aizu.ac.jp/~hamada

Today’s Topics

- **What is Entropy**
- **Information Source**
- **Measure of Information**
- **Self-Information**
- **Unit of Information**
- **Entropy**
- **Properties**

Entropy

Entropy of What?

- We need a **theory** of information and compression
- An abstract model that
  - Fits many situations
  - Is amenable to mathematical analysis
  - Stimulates predictions about what may be possible
  - Motivates search for new ideas about how to achieve what the theory says is possible

Entropy of What?

- In Simple words:

Entropy = Uncertainty

Or un-expected events.

Entropy

Low Entropy

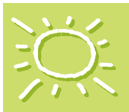
High Entropy

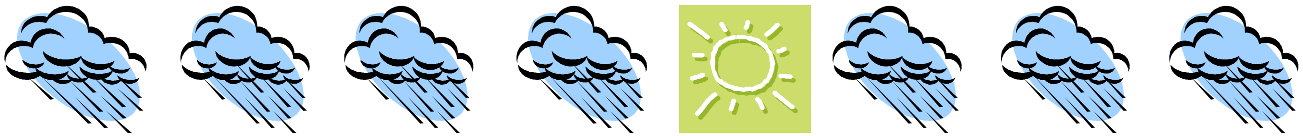
..the values (locations of soup) sampled entirely from within the soup bowl

..the values (locations of soup) unpredictable... almost uniformly sampled throughout the dining room

Low and High Information Content Messages

- The more frequent a message is, the less information it conveys when it occurs
- A weather forecast messages for Aizu in January for 8 days:

Sunny  **snow** 

Aizu: 

- "Snow" is a low information message and "Sunny" is a high information message

Information Theory

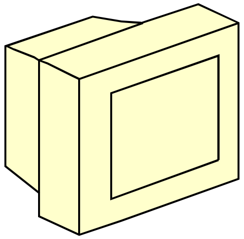
- The art of **quantifying** **and** **communicating** information

- Two Main Concepts
  - **ENTROPY**
  - **CAPACITY**

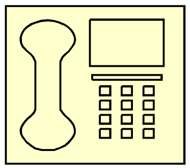
What is Entropy?

- A quantifiable means to describe the output of an information source

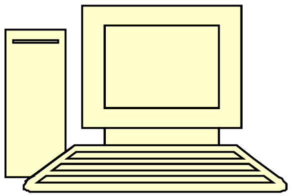
Video



Audio



Data



HOW COULD WE REPRESENT/COMPARE THE AMOUNT OF INFORMATION GENERATED AT THE OUTPUT OF THESE SOURCES?

So What's the Point?

- **The same information source, i.e., TV**

- In one scenario the information could be relayed in one sentence

- In the other scenario much more information needs to be communicated

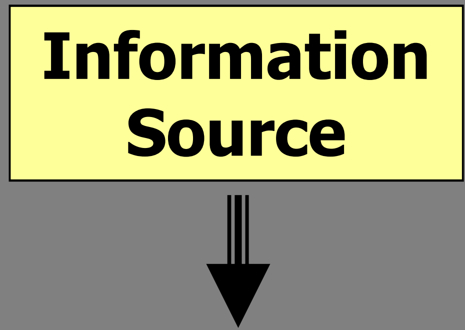
ENTROPY REFLECTS HOW MUCH INFORMATION IS THE SOURCE GENERATING IN EACH SCENARIO

What will you Learn from Source Coding?

- What is Entropy? (i.e., How could we quantify information)
- Fundamental Limits
  - The **minimum** number of bits per symbol to represent an information source
  - The **maximum** rate at which information transmission can take place over a channel

Information Source

- Assume
  - Information Source Generates a group of symbols from a given alphabet **S**
  - Symbols are independent (**Discrete Memoryless Source**)
- **Example**
  - The English Language
    - Alphabet **S**={a,b,c,…,z}
    - Each symbol has a probability: p_a, p_b , …p_z



$$S = \{s_0, s_1, \ldots, s_{K-1}\}$$

$$\Pr\{s_k\} = p_k, k = 0, 1, \ldots, K-1$$

$$\sum_{k=0}^{K-1} p_k = 1$$

Uncertainty and Measure of Information

- If $p_k=1$, $p_i=0$ i? k
  - There no uncertainty. The occurrence of the event does not correspond to any gain of information.
  - There is **no need** for communications because the receiver knows pretty much everything
- As p_k decreases,
  - The uncertainty increases
  - The reception of s_k corresponds to some gain in information. **BUT HOW MUCH?**

Measure of Information

- The amount of Information gained after observing s_k which has a probability p_k

$$I(s_k) = \log\left(\frac{1}{p_k}\right)$$

Self-Information = H(S) = lg(1/p)

- Greater frequency <==> Less information
- Extreme case: p = 1, H(S) = lg(1) = 0
- Why is this the right formula?
- 1/p is the average length of the gaps between recurrences of S

…..S…..S……………………..S…….S……S…..

a b c d

Average of a, b, c, d … = 1/p

Number of bits to specify a gap is about lg(1/p)

Unit of Information

- Depends on the BASE of the Logarithm

$$I(s_k) = \log_2\left(\frac{1}{p_k}\right)$$ **Bits**

$$I(s_k) = \log_e\left(\frac{1}{p_k}\right)$$ **Nats**

$$I(s_k) = \log_{10}\left(\frac{1}{p_k}\right)$$ **Hartleys**

Self-Information

Example

- If a symbol S has frequency p, its self-information is I(S) = lg(1/p) = -lg p.

S	A	B	C	D
p	.25	.25	.25	.25
I(S)	2	2	2	2
p	.7	.1	.1	.1
I(S)	.51	3.32	3.32	3.32

Properties of $I(s_k)$

1. $I(s_k)=0$ for $p_k=1$
2. $I(s_k)=0$ for $0=p_k =1$
3. $I(s_k)>I(s_i)$ for $p_k< p_i$
4. $I(s_k s_i)> I(s_k) + I(s_i)$ if s_k and s_i are independent

Entropy H(S)

- Entropy is the average information content of a source

$$H(S) = E\left[I(s_k)\right]$$

$$H(S) = \sum_{k=0}^{K-1} p_k \log_2\left(\frac{1}{p_k}\right)$$

Entropy

The larger the entropy of a distribution…

…the harder it is to predict

…the harder it is to compress it

…the less spiky the distribution

Entropy

Example

S	A	B	C	D	-? plgp
p	.25	.25	.25	.25	
-lgp	2	2	2	2	
-plgp	.5	.5	.5	.5	2
p	.7	.1	.1	.1	
-lgp	.51	3.32	3.32	3.32	
-plgp	.357	.332	.332	.332	1.353

What is the use of H(X)?

Shannon's first theorem (noiseless coding theorem)

For a memoryless discrete source X, its entropy H(X) defines the minimum average code length required to noiselessly code the source.

entropy

The **minimum** average number of binary digits needed

to

specify a source output (message) uniquely is called

"SOURCE ENTROPY"

Entropy of English

Example

- **Shannon Experiment**
  - given a sequence of characters
  - ask speaker of language to predict what the next character might be
  - record the number of guesses taken to get the right character
  - $H(\text{English}) = -1/n \sum p(\text{guess} = \text{character}) \log p(\text{guess} = \text{character})$
    - over all characters (letters and space)
    - n is 27

Example 3

- Calculate the Entropy of English language if
  1. All alphabet letters are equally probable
  2. For a, e, o, t $P\{s_k\}=0.1$

     For h, i, n, r, s $P\{s_k\}=0.07$

     For c, d, f, l, m, p, u, y $P\{s_k\}=0.02$

     For b, g, j, k, q, v, w, x, z $P\{s_k\}=0.01$

1. $H(S)=4.7$ bits
2. $H(S)=4.17$ bits

Some Properties of Entropy

- $0 = H(S) = \log_2(K)$
  1. $H(S)=0$ if and only if $p_k=1$ for some k and all the remaining probabilities are zero (NO UNCERTAINTY)
  2. $H(S)=\log_2(K)$ if and only if $p_k=1/K$ for all k Symbols of S are equiprobable

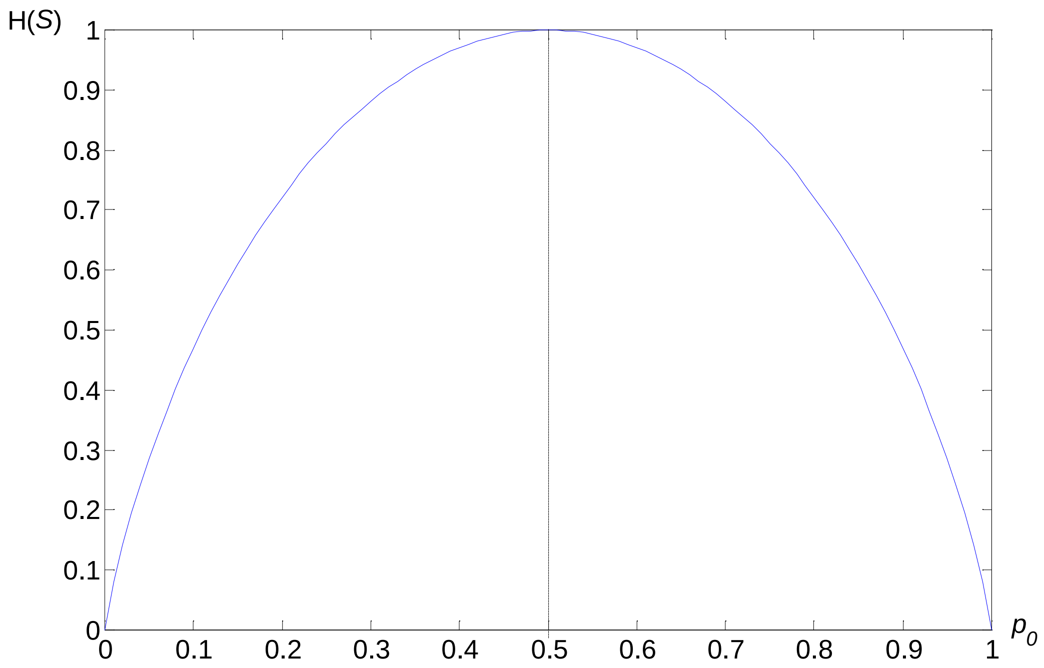
the "worst" we can do is

just assign $\log_2 M$ bits to each source output

Example 1: Entropy of a Binary Memoryless Source

- $S=\{s_0,s_1\}$, $\Pr\{s_0\}=p_0$, $\Pr\{s_1\}=p_1=1-p_0$

$$H(S)=-p_0\log_2(p_0)-(1-p_0)\log_2(1-p_0)$$



Entropy

Comments

Entropy is a measure of how much information is encoded in a message. Higher the entropy, higher the information content.

We could also say entropy is a measure of uncertainty in a message.

Information and uncertainty are equivalent concepts.

Entropy gives the actual number of bits of information contained in a message source.

Example: if the probability of the character \`e\` appearing in this slide is 1/16, then the information content of this character is 4 bits.

So the character string \`eeeee\` has a total of 20 bits (contrast this to using an 8-bit ASCII coding that could result in 40 bits to represent \`eeeee\`.