15-750:Algorithms in the Real World

Data Compression

Compression in the Real World

Ubiquitous usage. Examples:

- Data storage: file systems, large-scale storage systems (e.g. cloud storage)
- Communication
- Media: Video, audio, images
- Data structures: Graphs, indexes
- Newer: Neural network compression

Encoding/Decoding

"Message" refers to the data to be compressed

The encoder and decoder need to understand common compressed format.

Lossless vs. Lossy

Lossless: Input message = Output message

Lossy: Input message $\approx$ Output message

Quality of Compression:

For Lossless?

Runtime vs. Compression vs. Generality

For Lossy?

Loss metric (in addition to above)

How much can we compress?

Q: Can we (lossless) compress any kind of messages?

No!

For lossless compression, assuming all input messages are valid, if one string is compressed, some other must expand.

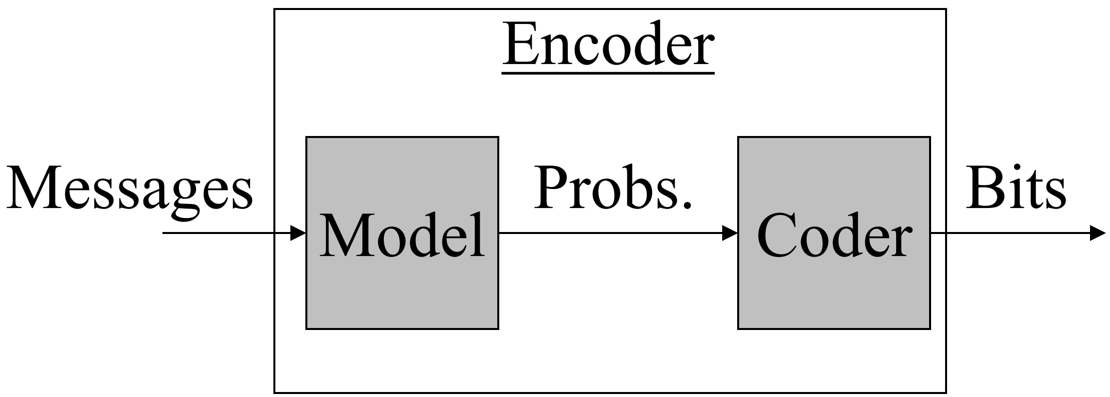
Q: So what we do need in order to be able to compress?

Can compress only if some messages are more likely than other.

That is, there needs to be **bias** in the probability distribution.

Model vs. Coder

To compress we need a bias on the probability of messages. The **model** determines this bias



Example models:

- Simple: Character counts, repeated strings
- Complex: Models of a human face

INFORMATION THEORY BASICS

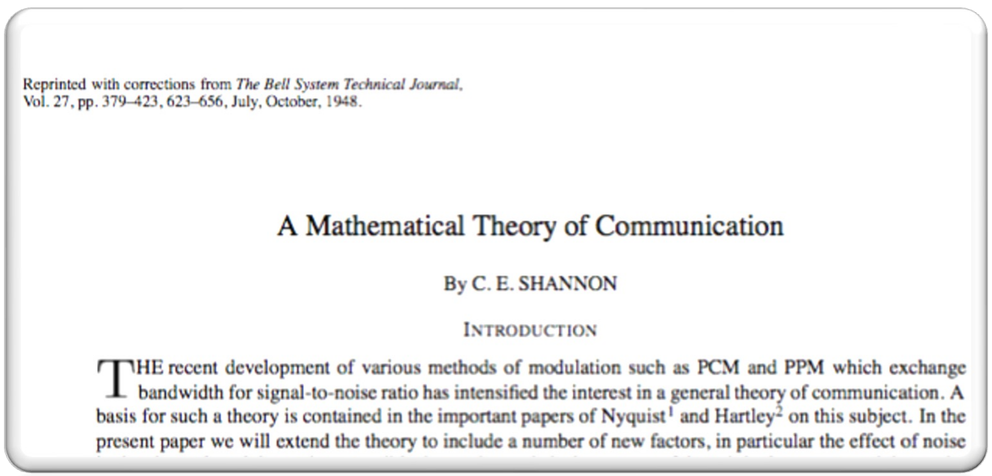
Information Theory

- Quantifies and investigates "information"
- Fundamental limits on representation and transmission of information
  - What's the minimum number of bits needed to **represent** data?
  - What's the minimum number of bits needed to **communicate** data?
  - What's the minimum number of bits needed to **secure** data?

Information Theory

Claude E. Shannon

- Landmark 1948 paper: mathematical framework
- Proposed and solved key questions
- Gave birth to information theory



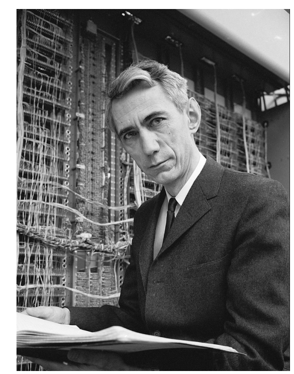
Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise



Information Theory

In the context of compression:

An interface between modeling and coding

Entropy

– A measure of information content

Suppose a message can take **n** values from $S = \{s_1, \ldots, s_n\}$ with a probability distribution ***p(s)***.

One of the n values will be chosen.

"How much choice" is involved? OR

"How much information is needed to convey the value chosen?

Entropy

Q: Should it depend on the values $\{s_1, \ldots, s_n\}$?

(e.g., American names vs. European names)

No.

Q: Should it depend on p(s)?

Yes.

If P(s_1)=1 and rest are all 0?

No choice. Entropy = 0

More the bias lower the entropy

Entropy

Shannon (1948 paper) lists key properties that an entropy function should satisfy and *shows that "log" is the only function*.

Specifically, $\log\left(\frac{1}{p(s)}\right)$

Intuition for the log function:

- When p(s) is low, entropy should be high
- Suppose two independent messages are being picked then entropy should add up

Entropy

For a set of messages S with probability $p(s)$, $s \in S$, the **self information** of s is:

$$i(s) = \log \frac{1}{p(s)} = -\log p(s)$$

Measured in **bits if the log is base 2**.

Entropy is the weighted average of self information.

$$H(S) = \sum_{s \in S} p(s) \log \frac{1}{p(s)}$$

Entropy Example

Binary random variable (i.e., taking two values) with probability p and 1-p

Denoted as $H_2(p)$:

<draw>

Highest entropy when equiprobable

(true for n >2 as well)

Entropy Example

$$p(S) = \{.25, .25, .25, .125, .125\}$$

$$H(S) = 3 \times .25 \log 4 + 2 \times .125 \log 8 = 2.25$$

$$p(S) = \{.5, .125, .125, .125, .125\}$$

$$H(S) = .5 \log 2 + 4 \times .125 \log 8 = 2$$

$$p(S) = \{.75, .0625, .0625, .0625, .0625\}$$

$$H(S) = .75 \log(4/3) + 4 \times .0625 \log 16 = 1.3$$

Conditional Entropy

Conditional entropy: Information content based on a context

The **conditional probability** *p(s|c)* is the probability of *s* in a context *c*.

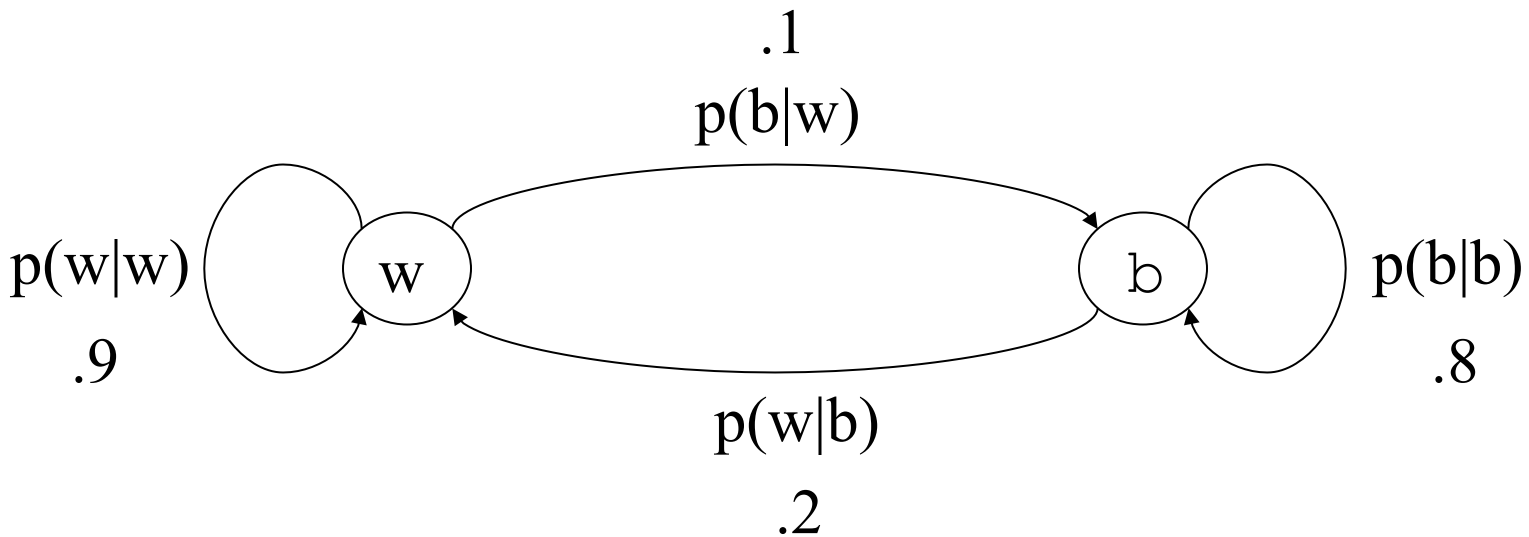
The **conditional entropy** is the weighted average of the conditional self information

$$H(S \mid C) = \sum_{c \in C} \left(p(c) \sum_{s \in S} p(s \mid c) \log \frac{1}{p(s \mid c)} \right)$$

Types of “sources”

- Sources generate the messages (to be compressed)
- Sources can be modelled in multiple ways
- Independent and identically distributed (i.i.d) source
  - Prob. of each msg is independent of the previous msg
- Markov source
  - message sequence follows a Markov model (specifically Discrete Time Markov Chain, aka DTMC)

Example of a Markov Chain



Shannon's experiment

Asked people to predict the next character given the whole previous text. He used these as conditional probabilities to estimate the entropy of the English Language.

The number of guesses required for right answer:

# of guesses	1	2	3	4	5	> 5
Probability	.79	.08	.03	.02	.02	.05

From the experiment

H(English) = .6 - 1.3

In comparison, ASCII uses 7 bits, Unicode and other representations use 8 or even higher