# Information Theory and Networks

## Lecture 5: Entropy

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/
Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

September 18, 2013

# Part I

# Entropy

You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.

*John von Neumann, Suggesting to Claude Shannon a name for his new uncertainty function, as quoted in Scientific American Vol. 225 No. 3, (1971), p. 180*

# Section 1

# Entropy: definitions

# Entropy

Entropy will be our measure of uncertainty.

Let $X$ be a discrete random variable with alphabet $\Omega$ and PMF $p(x)$. The only definition of Entropy that satisfies all of our axioms is

## Definition (Entropy)

(Shannon) entropy is defined to be

$$H(X) = -\sum_{x \in \Omega} p(x) \log_2 p(x),$$

We might also write $H(\mathbf{p})$ for the same quantity.

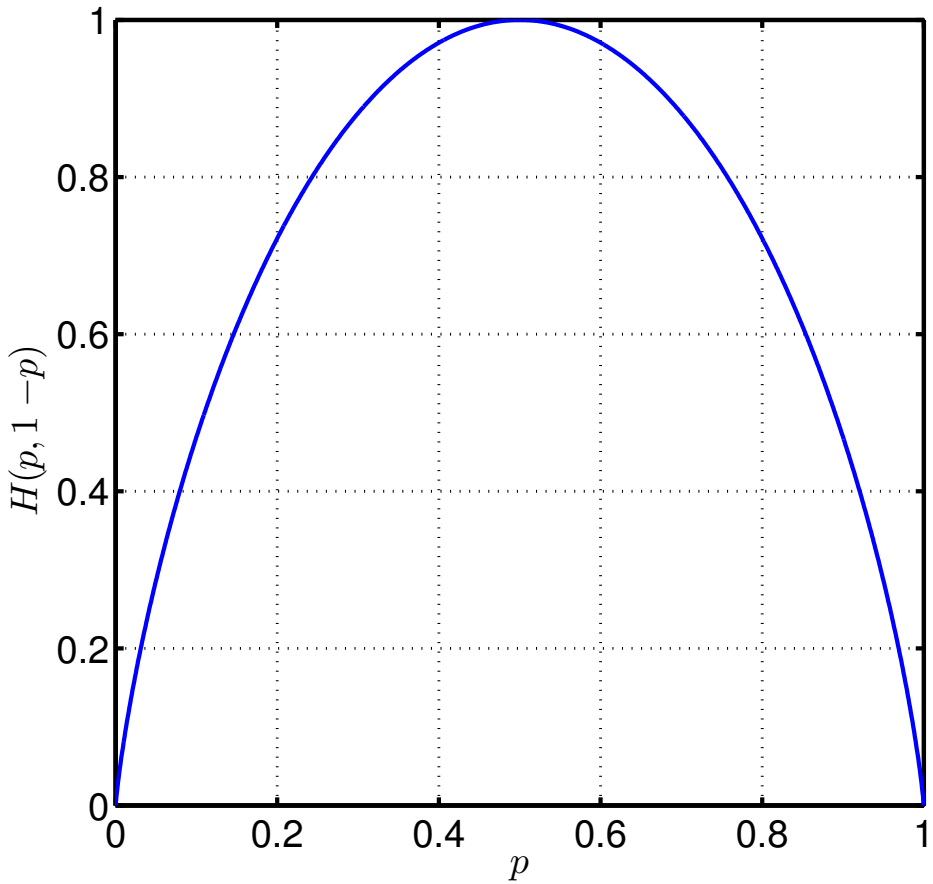
# Entropy Example 1: Bernoulli RV

For a Bernoulli random variable with: $\Omega = \{0, 1\}$

$$p(1) = p, \text{ and } p(0) = 1 - p = q$$

We get

$$H(p, 1-p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



# Entropy Example 2: [CT91, p.14]

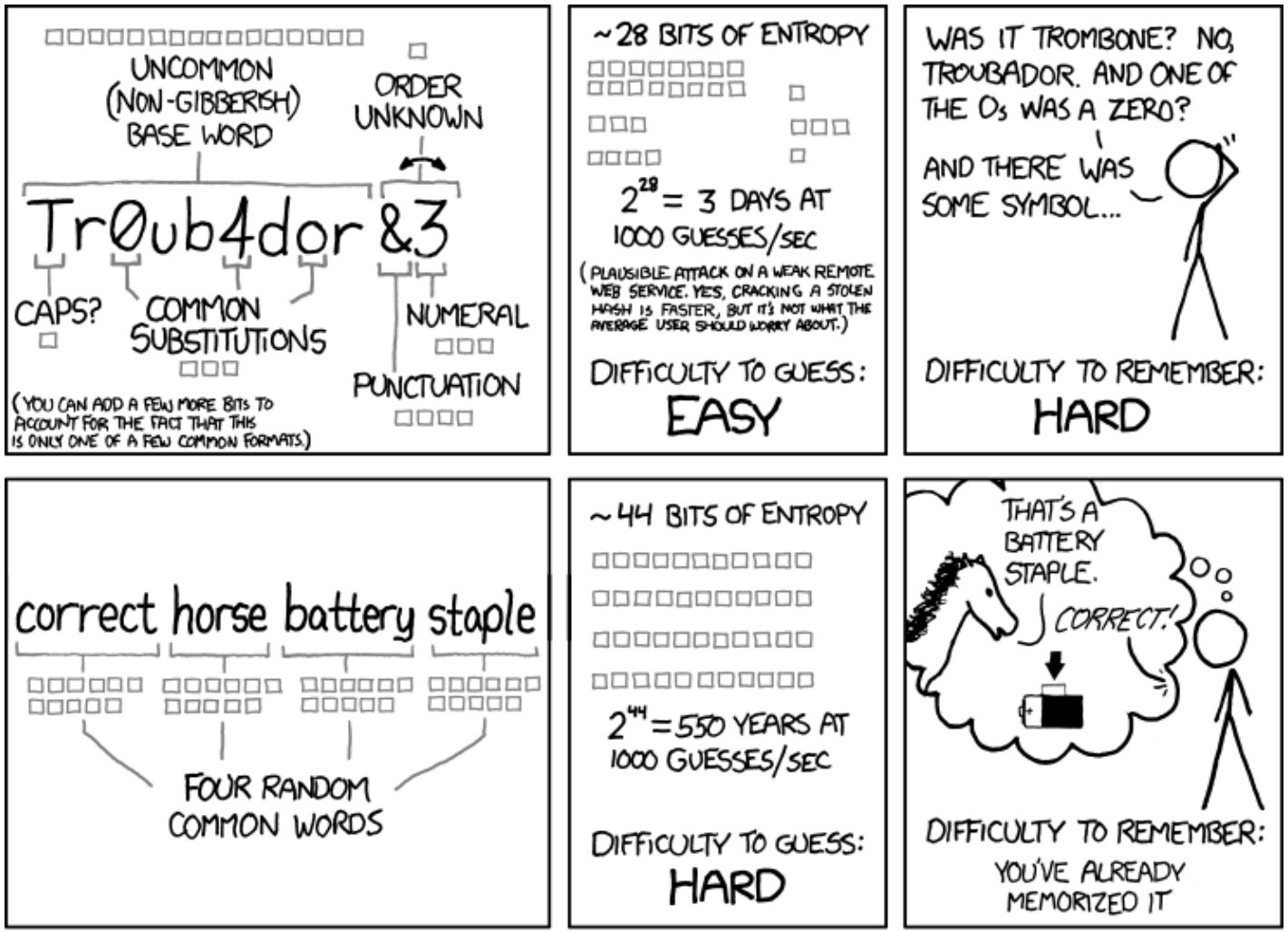
Take symbols $\Omega = \{a, b, c, d\}$, with probabilities

$$X = \begin{cases} a, & \text{with probability } 1/2, \\ b, & \text{with probability } 1/4, \\ c, & \text{with probability } 1/8, \\ d, & \text{with probability } 1/8, \end{cases}$$

Then

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4} \text{ bits.}$$

# Entropy Example 3: https://xkcd.com/936/



# Joint Entropy

## Definition (Joint Entropy)

Given two discrete RVs $X$ and $Y$ with joint distribution $p(x, y)$ the joint entropy is defined to be

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log_2 p(x, y),$$

This shouldn't be surprising, as its just the same definition on the alphabet $(x, y)$.

# Conditional Entropy

## Definition (Conditional Entropy)

Given two discrete RVs $X$ and $Y$ with joint distribution $p(x, y)$ the conditional entropy of $Y$ given $X$ is defined to be

$$\begin{aligned} H(Y|X) &= -E\left[\log p(Y|X)\right] \\ &= -\sum_x \sum_y p(x,y) \log p(y|x) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= -\sum_x p(x) H(Y|X=x). \end{aligned}$$

where $p(x)$ is the marginal distribution of $X$.

# Conditional Entropy: examples

- Perfect dependence: $Y = f(X)$, so given $X$ there is no uncertainty about $Y$, then
$$H(Y|X) = 0$$
- Independence: $p(y|x) = p(y)$, so
$$H(Y|X) = H(Y),$$
so uncertainty of $Y$ is unchanged by knowledge of $X$.

# Relative Entropy

Relative entropy is an asymmetric measure of

- the "distance" between two distributions
- inefficiency of assuming $q$ when $p$ is true

## Definition (Relative entropy)

The relative entropy or Kullback-Leibler divergence is a measure of the distance from PMF $p(x)$ to PMF $q(x)$ and is defined by

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \left[ \log \frac{p(X)}{q(X)} \right].$$

# Relative Entropy Example: [CT91, p.17]

Let $(X, Y)$ have the following values

| | | X | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | 1 | 1/8 | 1/16 | 1/32 | 1/32 |
| Y | 2 | 1/16 | 1/8 | 1/32 | 1/32 |
| | 3 | 1/16 | 1/16 | 1/16 | 1/16 |
| | 4 | 1/4 | 0 | 0 | 0 |

The marginal distributions are

$$\begin{aligned} p(X) &= (1/2, 1/4, 1/8, 1/8) \\ p(Y) &= (1/4, 1/4, 1/4, 1/4) \end{aligned}$$

so $H(X) = 7/4$ and $H(Y) = 2$ bits.

# More Complex Entropy Example: [CT91, p.17]

Joint entropy

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 p(x,y) = 27/8 \text{ bits}$$

Conditional entropies

$$H(Y|X) = -\sum_x p(x) H(Y|X=x) = 13/8 \text{ bits}$$

$$H(X|Y) = -\sum_y p(y) H(X|Y=y) = 11/8 \text{ bits}$$

Section 2

# A Brief History of Information Theory

# Entropy

- First though about in the context of physics: statistical mechanics
  - Ludwig Boltzmann, 1872
  - J. Willard Gibbs, 1878
- More on that connection later

# Context

Telephone and Telegraphy grows massively

- Invented 1753?
- Concrete idea Samuel Soemmering in 1809
- Morse and Vail (not just code) 1835
- First serious demonstrator: Washington to Baltimore, a distance of 40 miles, was completed in 1844
  - The first message, composed by Annie Ellsworth, the young daughter of Morse's friend was "What hath God wrought?".
- First undersea cable Sept 1851 across English channel
- 1865 there were 83,000 miles of wire in the USA.
- First transatlantic line 1866
- Society of Telegraph Engineers was founded in 1871
- Todd's telegraphs importance to Australia 1872
- 1882 Bell Lab is created
- 1904 photograph transmitted by wire in Germany
- 1907, the US alone had around 3 million miles of telephone and telegraph wires
- The figure was 67.8 million miles by 1925

# Information

- 1924, Harry Nyquist starts formalising transmission capacities
  - "intelligence" and the speed it can be transmitted
- 1928, Ralph Hartley introduces Hartley Information, as log of number of possible messages (or log of alphabet size)
- 1940, Alan Turing introduces the deciban in relationship to finding cypher settings

# Shannon and Information Theory

- By the 1940, AT&T/Bell had
  - nearly 100 million miles of telephone and telegraph cable
  - 280,000 employees
  - 80 million daily telephone calls
  - \$1.2 billion revenue
- Along comes Shannon (joins Bell Labs in 1942)
  - worked for AT&T/Bell
  - influenced by
    - ★ at princeton: Hermann Weyl, von Neumann, Einstein, Gödel
    - ★ doing crypto during war: Alan Turing
    - ★ MIT: Vannevar Bush
  - thought about TV, genes, cryptography, ...
- Newton made force into a quantity with units, Shannon made information into a quantity with units (bits)

# Later developments

- 1944, Shannon's theory mainly complete, but main publication in 1948
- 1947, Hamming codes
- 1949, Fano proves some basic results
- 1951, Relative entropy by Kullback and Leibler
- 1950s onwards various people used the ideas for coding (Huffman, Reed, Muller, Solomon, Gallager, Viterbi, ...)
- 1957, Jayne relates information theory back to statistics and physics

# Today

More important than ever

- Mp3, video, voice, ...
- Internet
- Digital TV and radio
- Bio*informatics*
- Google and Big Data

Over 1.5 billion miles of "telephone" wire are said now to be strung across the U.S.

Anything you do as a scientist or mathematician will be influenced by information theory, whether you know it or not.

# Further reading I



Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.


James Gleick, *The information: a history, a theory, a flood*, Fourth Estate, 2011.