Lecture 9
A Brief Intro to Information Theory

- Motivation
- Entropy and Mutual Information

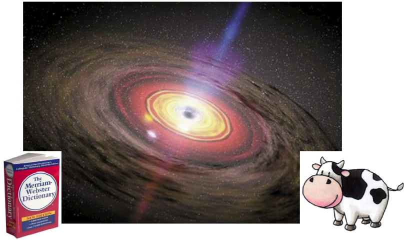
Contents

A Thought Experiment

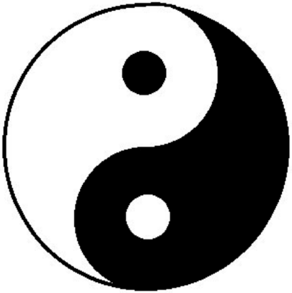
Throw a cow or a dictionary into a black hole,

which has higher information loss?

\- Tom Cover



How to quantify information?

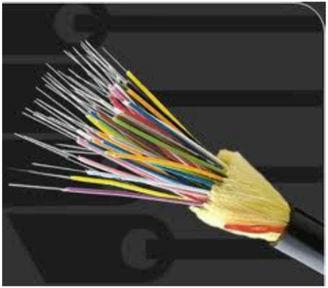


Small information content

Large information content

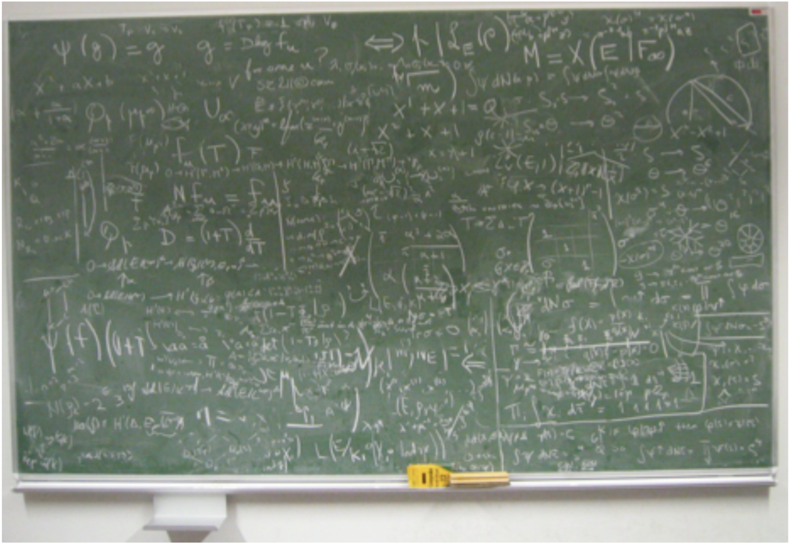
What is the fundamental limit of data transfer rate?

WiFi: data rate ~ Mbit/s

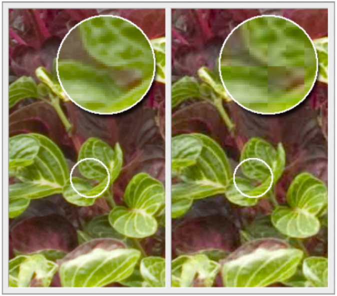


Fiber Optics: data rate ~ Tbit/s

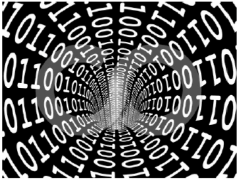
Some people think information theory (IT) is about...



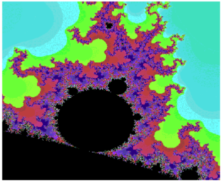
But IT is also about these...



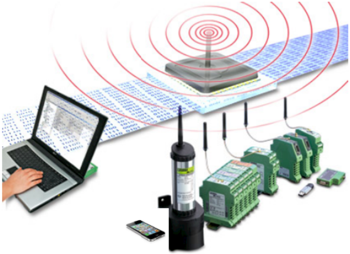
Data Compression



Coding

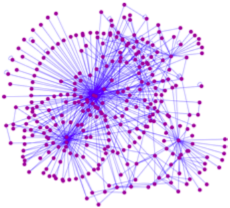


Computation: Kolmogorov Complexity

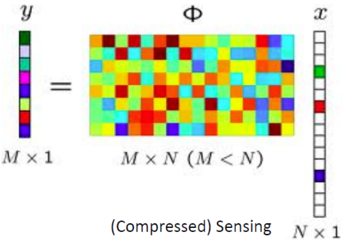


Data Communication

And even these...

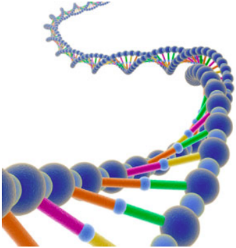


Network



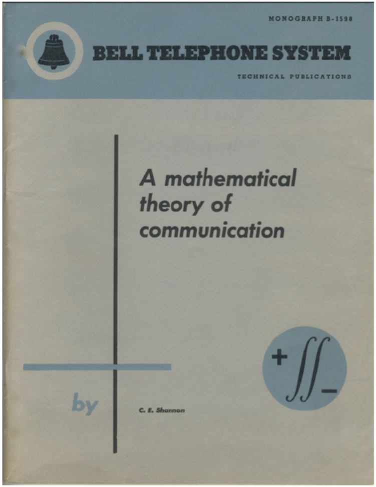
(Compressed) Sensing

Investment, gambling

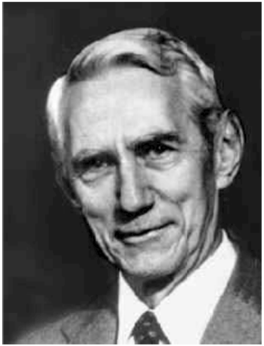


Bioinformatics

Where IT all begins...



1948, Bell Sys. Tech. Journal



Shannon, 1916 - 2001

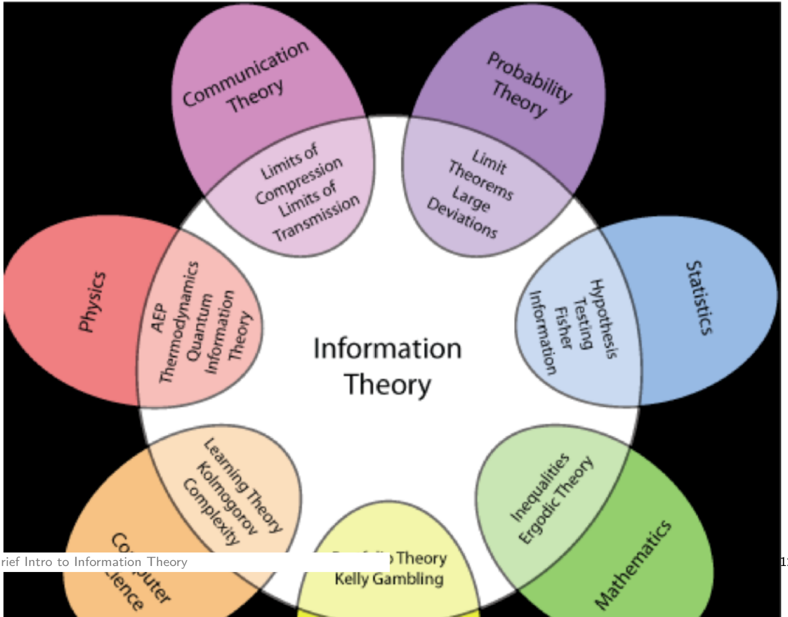
Information Theory

- Shannon's information theory deals with limits on data compression (source coding) and reliable data transmission (channel coding)
  - How much can data be compressed?
  - How fast can data be reliably transmitted over a noisy channel?
- Two basic "point-to-point" communication theorems (Shannon 1948)
  - **Source coding theorem**: the minimum rate at which data can be *compressed losslessly* is the *entropy rate* of the source
  - **Channel coding theorem**: The maximum rate at which data can be *reliably transmitted* is the *channel capacity* of the channel

Extensions and Applications

- Since Shannon's 1948 paper, many extensions
  - Rate distortion theory
  - Source coding and channel capacity for more complex sources
  - Capacity for more complex channels (multiuser networks)
- Information theory was considered (by most) an esoteric theory with no apparent relation to the "real world"
- Recently, advances in technology (algorithms, hardware, software) today there are practical schemes for
  - data compression
  - transmission and modulation
  - error correcting coding
  - compressed sensing techniques
  - information security ...

IT encompasses many fields



In this class we will cover the basics

- **Nuts and Bolts**
  - Entropy: uncertainty of a single random variable

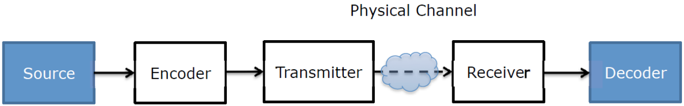
$$H(X) = -\sum_x p(x) \log_2 p(x) \text{ (bits)}$$

  - Conditional Entropy: $H(X|Y)$
  - Mutual information: reduction in uncertainty due to another random variable

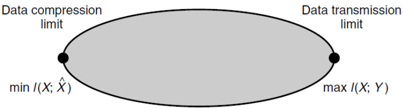
$$I(X;Y) = H(X) - H(X|Y)$$

  - Channel capacity $C = \max_{p(x)} I(X;Y)$
  - Relative entropy: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

Fundamental Limits



- – Data compression limit (lossless source coding)
  – Data transmission limit (channel capacity)
  – Tradeoff between rate and distortion (lossy compression)



Important Functionals

- Upper case $X, Y, \ldots$ refer to random variables
- $\mathcal{X}, \mathcal{Y}$ alphabet of random variables
- $p(x) = P(X = x)$
- $p(x, y) = P(X = x, Y = y)$
- Probability density function $f(x)$

Expectation and Variance

- **Expectation**: $\mu = \mathbb{E}\{X\} = \sum xp(x)$
- Why is this of particular interest? It appears in Law of Large Number (LLN): If x_n independent and identically distributed,

$$\frac{1}{N}\sum_{n=1}^{N} x_n \to \mathbb{E}\{X\}, \text{ w.p.1}$$

- **Variance**: $\sigma^2 = \mathbb{E}\{(X-\mu)^2\} = \mathbb{E}\{X^2\} - \mu^2$
- Why is this of particular interest? It appears in Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{N\sigma^2}}\sum_{n=1}^{N}(x_n - \mu) \to \mathcal{N}(0,1)$$

Information theory: is it all about theory?

Yes and No.

Yes, it's theory

- Yes, it's theory. We will see many proofs. But it's also in preparation for other subjects
  - Coding theory (Prof. R. Calderbank)
  - Wireless communications
  - Compressed sensing
  - Stochastic network
  - Many proof ideas come in handy in other areas of research

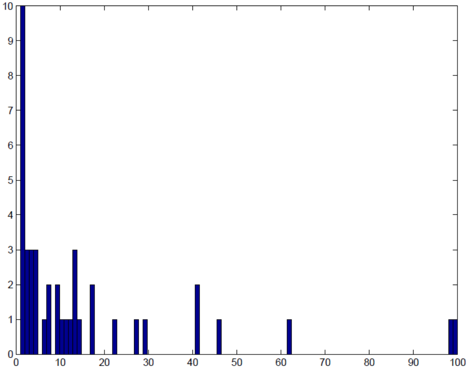
No, it's practical too

- No. Hopefully you will walk out of this classroom understanding
  - Basic concepts people talk on the streets: entropy, mutual information ...
  - Channel capacity - all wireless guys should know
  - Huffman code (the optimal lossless code)
  - Hamming code (commonly used single error correction code)
  - "Water-filling" - power allocation in all communication systems
  - Rate-distortion function - if you want to talk with data compression guy

Contents

The winner is:

Eunsu Ryu, with number 6



A strategy to win the game?

The winner is:

Which horse won?

Uncertainty measure

- Let X be a random variable taking on a finite number M of different values $x_1, \cdots, x_M$
- What is X: English letter in a file, last digit of Dow-Jones index, result of coin tossing, password
- With probability $p_1, \cdots, p_M$, $p_i > 0$, $\sum_{i=1}^{M} p_i = 1$
- Question: what is the uncertainty associated with X?
- Intuitively: a few properties that an uncertainty measure should satisfy
- It should not depend on the way we choose to label the alphabet

Desired properties

- It is a function of $p_1, \cdots, p_M$
- Let this uncertainty measure be

$$H(p_1, \cdots, p_M)$$

- **Monotonicity**. Let $f(M) = H(1/M, \cdots, 1/M)$. If $M < M'$, then

$$f(M) < f(M')$$

- Picking one person randomly from the classroom should result less possibility than picking a person randomly from the US.

Desired properties (continued)

- **Additivity**. Two independent RV X and Y, each uniformly distributed, alphabet size M and L. The uncertainty for the pair (X, Y), is ML. However, due to independence, when X is revealed, the uncertainty in Y should not be affected. This means

$$f(ML) - f(M) = f(L)$$

- **Grouping rule** (Problem 2.27 in Text). Dividing the outcomes into two, randomly choose one group, and then randomly pick an element from one group, does not change the number of possible outcomes.

Entropy

- The only function that satisfies the requirements is the entropy function

$$H(p_1, \cdots, p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

- General definition of entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \text{ bits}$$

- $0 \log 0 = 0$

Understanding Entropy

- Uncertainty in a single random variable
- Can also be written as:

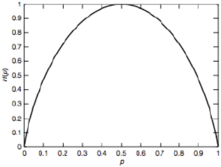
$$H(X) = \mathbb{E}\left\{\log \frac{1}{p(X)}\right\}$$

- Intuition: $H = \log(\texttt{\#of outcomes/states})$
- Entropy is a functional of $p(x)$
- Entropy is a lower bound on the number of bits need to represent a RV. E.g.: a RV that has uniform distribution over 32 outcomes

Properties of entropy

- $H(X) \geq 0$
- Definition, for Bernoulli random variable, $X = 1$ w.p. p, $X = 0$ w.p. $1 - p$

$$H(p) = -p \log p - (1-p)\log(1-p)$$



- **Concave**
- Maximizes at $p = 1/2$
- Example: how to ask questions?

Joint entropy

- Extend the notion to a pair of discrete RVs (X, Y)
- Nothing new: can be considered as a single vector-valued RV
- Useful to measure dependence of two random variables

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

$$H(X, Y) = -\mathbb{E} \log p(X, Y)$$

Conditional Entropy

- Conditional entropy: entropy of a RV given another RV. If $(X, Y) \sim p(x, y)$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

- Various ways of writing this

Chain rule for entropy

- Entropy of a pair of RVs = entropy of one + conditional entropy of the other:

$$H(X,Y) = H(X) + H(Y|X)$$

- Proof:
- $H(Y|X) \neq H(X|Y)$
- $H(X) - H(X|Y) = H(Y) - H(Y|X)$

Relative entropy

- Measure of distance between two distributions

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- Also known as Kullback-Leibler distance in statistics: expected log-likelihood ratio
- A measure of inefficiency of assuming that distribution is q when the true distribution is p
- If we use distribution is q to construct code, we need $H(p) + D(p||q)$ bits on average to describe the RV

Mutual information

- Measure of the amount of information that one RV contains about another RV

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$$

- Reduction in the uncertainty of one random variable due to the knowledge of the other
- Relationship between entropy and mutual information

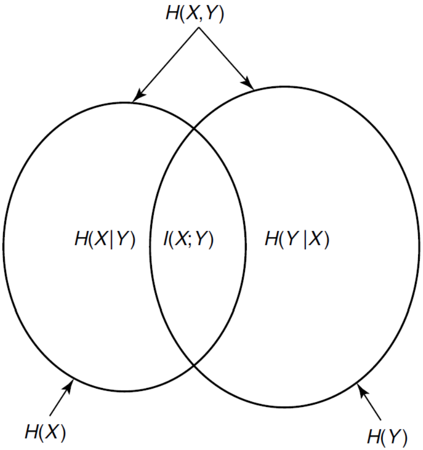
$$I(X;Y) = H(Y) - H(Y|X)$$

- Proof:

Mutual information properties

- $I(X;Y) = H(Y) - H(Y|X)$
- $H(X,Y) = H(X)+H(Y|X) \rightarrow I(X;Y) = H(X)+H(Y)-H(X,Y)$
- $I(X;X) = H(X) - H(X|X) = H(X)$ Entropy is "self-information"
- Example: calculating mutual information

Venn diagram



$I(X;Y)$ is the intersection of information in X with information in Y

Example: Blood type and skin cancer risk

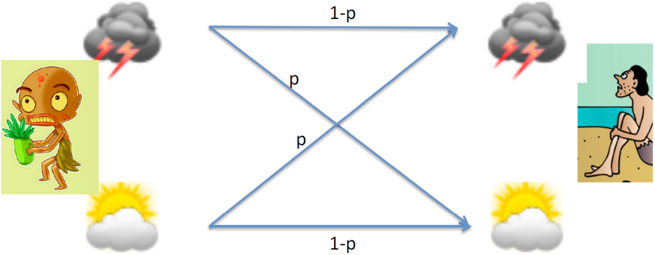
X: blood type

Y: chance for skin cancer

	A	B	AB	O
Very Low	1/8	1/16	1/32	1/32
Low	1/16	1/8	1/32	1/32
Medium	1/16	1/16	1/16	1/16
High	1/4	0	0	0

- X: marginal (1/2, 1/4, 1/8, 1/8)
- Y: marginal (1/4, 1/4, 1/4, 1/4)
- $H(X) = 7/4$ bits $\quad H(Y) = 2$ bits
- Conditional entropy: $H(X|Y) = 11/8$ bits, $H(Y|X) = 13/8$ bits
- $H(Y|X) \neq H(X|Y)$
- Mutual information: $I(X;Y) = H(X) - H(X|Y) = 0.375$ bit

Example: Binary Symmetric Channel

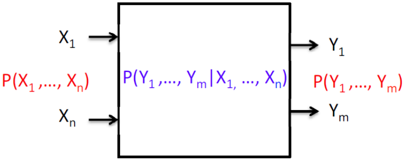


Summary

Entropy

H(X)

Mutual Information



$I(X_1,\ldots, X_n; Y_1, \ldots, Y_m)$