# Introduction to Information Theory

**David Feldman**

July 2004

College of the Atlantic
and
Santa Fe Institute

**dave@hornacek.coa.edu**

**http://hornacek.coa.edu/dave/**

## Information Theory

- Originally developed by Shannon in 1948 as he was figuring out how to efficiently transmit communication signals over a possibly noisy communication channel.
- I am not so much interested in its original uses in communication theory, but in its development and application as a broadly applicable tool for describing probability distributions.
- Information theory lets us ask and answer questions such as:
  1. How random is a sequence of measurements?
  2. How much memory is needed to store the outcome of measurements?
  3. How much information does one measurement tell us about another?

## Some Info Theory References


1. T.M. Cover and J.A. Thomas, Elements of Information Theory. John Wiley & Sons, Inc., 1991. By far the best information theory text around.

2. C.E. Shannon and W. Weaver. The Mathematical Theory of Communication. University of Illinois Press. 1962. Shannon's original paper and some additional commentary. Very readable.

3. J.P. Crutchfield and D.P. Feldman, "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence." *Chaos* **15**:25–53. 2003.

4. Tom Schneider. Information Theory Primer, `http://www.lecb.ncifcrf.gov/~toms/paper/primer/`. Brief narrative at an elementary level. Includes a review or logarithms.

5. Entropy on the World Wide Web. `http://www.math.psu.edu/gunesch/entropy.html`. A huge list of links. Info theory applied to many different disciplines.

6. D.P. Feldman. A Brief Tutorial on: Information Theory, Excess Entropy and Statistical Complexity: Discovering and Quantifying Statistical Structure. `http://hornacek.coa.edu/dave/Tutorial/index.html`.

## Notation for Probabilities

Information theory is concerned with probabilities. We first fix some notation.

- $X$ is a random variable. The variable $X$ may take values $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite set.
- likewise $Y$ is a random variable, $Y = y \in \mathcal{Y}$.
- The probability that $X$ takes on the particular value $x$ is $\Pr(X = x)$, or just $\Pr(x)$.
- Probability of $x$ and $y$ occurring: $\Pr(X = x, Y = y)$, or $\Pr(x, y)$
- Probability of $x$, given that $y$ has occurred: $\Pr(X = x|Y = y)$ or $\Pr(x|y)$

Example: A fair coin. The random variable $X$ (the coin) takes on values in the set $\mathcal{X} = \{h, t\}$.

$\Pr(X = h) = 1/2$, or $\Pr(h) = 1/2$.

## Different amounts of uncertainty?

- Anytime we describe a situation with probabilities, it's because we're uncertain of the outcome.
- However, some probability distributions indicate more uncertainty than others.
- We seek a function $H[X]$ that measures the amount of uncertainty associated with outcomes of the random variable $X$.
- What properties should such an uncertainty function have?
  1. Maximized when the distribution over $X$ is uniform.
  2. Continuous function of the probabilities of the different outcomes of $X$
  3. Independent of the way in which we might group probabilities.

## Entropy of a Single Variable

The requirements on the previous page *uniquely* determine $H[X]$, up to a multiplicative constant.

The Shannon entropy of a random variable $X$ is given by:

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2(\Pr(x)) \ . \tag{1}$$

Using base-2 logs gives us units of *bits*.

**Examples**

- **Fair Coin:** $\Pr(h) = \frac{1}{2}, \Pr(t) = \frac{1}{2}$.
  $H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$ bit.
- **Biased Coin:** $\Pr(h) = 0.6, \Pr(t) = 0.4$.
  $H = -0.6\log_2 0.6 - 0.4\log_2 0.4 = 0.971$ bits.
- **More Biased Coin:** $\Pr(h) = 0.9, \Pr(t) = 0.1$.
  $H = -0.9\log_2 0.9 - 0.1\log_2 0.1 = 0.469$ bits.
- **Totally Biased Coin:** $\Pr(h) = 1.0, \Pr(t) = 0.0$.
  $H = -1.0\log_2 1.0 - 0.0\log_2 0.0 = 0.0$ bits.

We now consider various interpretations for the entropy.

## Average Surprise

- $-\log_2 x$ may be viewed as the *surprise* associated with the outcome $x$.
- Thus, $H[X]$ is the average, or expected value, of the surprise:

$$H[X] = \sum_x [-\log_2 x] \Pr(x) \ .$$

- The more surprised you are about a measurement, the more informative it is.
- The greater $H[X]$, the more informative, on average, a measurement of $X$ is.

## Difficulty of Guessing

For the next few slides, we'll focus on two examples.

1. A random variable $X$ with four equally likely outcomes: $\Pr(a) = \Pr(b) = \Pr(c) = \Pr(d) = \frac{1}{4}$.
2. A random variable $Y$ with four outcomes: $\Pr(\alpha) = \frac{1}{2}$, $\Pr(\beta) = \frac{1}{4}$, $\Pr(\gamma) = \frac{1}{8}$, $\Pr(\delta) = \frac{1}{8}$.

**What is the optimal strategy for guessing (via yes-no questions) the outcome of a random variable?**

- In general, try to divide the probability in half with each guess.
- Example: Guessing $X$:
  1. "is $X$ equal to $a$ or $b$?"
  2. If yes, "is $X = a$?" If no, "is $X = c$?"
- Using this strategy, it will always take $2$ guesses.
- $H[X] = 2$. Coincidence???

## Guessing games, continued

What's the best strategy for guessing $Y$?

$$\Pr(\alpha) = \frac{1}{2}, \Pr(\beta) = \frac{1}{4}, \Pr(\gamma) = \frac{1}{8}, \Pr(\delta) = \frac{1}{8}.$$

1. Is it $\alpha$? If yes, then done, if no:
2. Is it $\beta$? If yes, then done, if no:
3. Is it $\gamma$? Either answer, done.

Ave # of guesses = $\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) = 1.75$.

Not coincidentally, $H[Y] = 1.75$!!

**General result: Average number of yes-no questions needed to guess the outcome of $X$ is between $H[X]$ and $H[X] + 1$.**

- This is consistent with the interpretation of $H$ as uncertainty.
- If the probability is concentrated more on some outcomes than others, we can exploit this regularity to make more efficient guesses.

## Coding

- A *code* is a mapping from a set of symbols to another set of symbols.
- Here, we are interested in a code for the possible outcomes of a random variable that is as short as possible while still being decodable.
- Strategy: use short code words for the more common occurrences of $X$.
- This is identical to the strategy for guessing outcomes.

Example: Optimal binary code for $Y$:

$$\alpha \longrightarrow 1\ , \quad \beta \longrightarrow 01$$
$$\gamma \longrightarrow 001\ , \quad \delta \longrightarrow 000$$

Note: This code is unambiguously decodable:

$$0110010000000101 = \beta\alpha\gamma\delta\delta\gamma\gamma$$

This type of code is called an *instantaneous* code.

## Coding, continued

**General Result: Average number of bits in optimal binary code for $X$ is between $H[X]$ and $H[X] + 1$.**

This result is known as Shannon's noiseless source coding theorem or Shannon's first theorem.

- Thus, $H[X]$ is the average memory, in bits, needed to store outcomes of the random variable $X$.

## Summary of interpretations of entropy

- $H[X]$ is *the* measure of uncertainty associated with the distribution of $X$.
- Requiring $H$ to be a continuous function of the distribution, maximized by the uniform distribution, and independent of the manner in which subsets of events are grouped, uniquely determines $H$.
- $H[X]$ is the expectation value of the surprise, $-\log_2 \Pr(x)$.
- $H[X] \leq$ Average number of yes-no questions needed to guess the outcome of $X \leq H[X] + 1$.
- $H[X] \leq$ Average number of bits in optimal binary code for $X \leq H[X] + 1$.
- $H[X] = \lim N \to \infty \frac{1}{N} \times$ average length of optimal binary code of $N$ copies of $X$.

## Joint and Conditional Entropies

### Joint Entropy

- $H[X,Y] \equiv$
  $-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2(\Pr(x,y))$
- $H[X,Y]$ is the uncertainty associated with the outcomes of $X$ **and** $Y$.

### Conditional Entropy

- $H[X|Y] \equiv$
  $-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2 \Pr(x|y)$ .
- $H[X|Y]$ is the average uncertainty of $X$ given that $Y$ is known.

### Relationships

- $H[X,Y] = H[X] + H[Y|X]$
- $H[Y|X] = H[X,Y] - H[X]$
- $H[Y|X] \neq H[X|Y]$

# Mutual Information

## Definition

- $I[X;Y] = H[X] - H[X|Y]$
- $I[X;Y]$ is the average reduction in uncertainty of $X$ given knowledge of $Y$.

## Relationships

- $I[X;Y] = H[X] - H[X|Y]$
- $I[X;Y] = H[Y] - H[Y|X]$
- $I[X;Y] = H[Y] + H[X] - H[X,Y]$
- $I[X;Y] = I[Y;X]$

## Example 1

Two independent, fair coins, $C_1$ and $C_2$.

| $C_1$ | $C_2$ | |
|---|---|---|
| | $h$ | $t$ |
| $h$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| $t$ | $\frac{1}{4}$ | $\frac{1}{4}$ |

- $H[C_1] = 1$ and $H[C_2] = 1$.
- $H[C_1, C_2] = 2$.
- $H[C_1|C_2] = 1$. Even if you know what $C_2$ is, you're still uncertain about $C_1$.
- $I[C_1; C_2] = 0$. Knowing $C_1$ does not reduce your uncertainty of $C_2$ at all.
- $C_1$ carries no information about $C_2$.

## Example 2

Weather (rain or sun) yesterday $W_0$ and weather today $W_1$.

| | $W_1$ | |
|---|---|---|
| $W_0$ | $r$ | $s$ |
| $r$ | $\frac{5}{8}$ | $\frac{1}{8}$ |
| $s$ | $\frac{1}{8}$ | $\frac{1}{8}$ |

- $H[W_0] = 0.811$ and $H[W_1] = 0.811$.
- $H[W_0, W_1] = 1.549$.
- Note that $H[W_0, W_1] \neq H[W_0] + H[W_1]$.
- $H[W_1|W_0] = 0.738$.
- $I[W_0; W_1] = 0.074$. Knowing the weather yesterday, $W_0$, reduces your uncertainty about the weather today $W_1$.
- $W_0$ carries $0.074$ bits of information about $W_1$.

Note: The above statistics are consistent with the perfectly periodic pattern: $\cdots rrrrrrssrrrrrrssrrrrrrss \cdots$.

How could we detect if this was the actual pattern?

## Application: Maximum Entropy

- A common technique in statistical inference is the **maximum entropy method**.
- Suppose we know a number of average properties of a random variable. We want to know what distribution the random variable comes from.
- This is an underspecified problem. What to do?
- Choose the distribution that maximizes the entropy while still yielding the correct average values.
- This is usually accomplished by using Lagrange multipliers to perform a constrained maximization.
- The justification for the maximum entropy method is that it assumes no information beyond what is already known in the form of the average values.
- Another application: In other settings in which one wants to design a maximally predictive model, one often adjusts parameters to maximize the mutual information between input variables and those variables that are to be predicted.