# Information Theory Definitions and Facts

The information conveyed is the uncertainty removed which is also the expected surprise yielded. If S is the set of all possible outcomes for a random variable $x$, then the entropy must be

Entropy $\equiv$ Info conveyed $\equiv$ uncertainty removed $\equiv$ E (surprise yielded) $\equiv$

$$H(X) = E[\text{surprise\_at}(X)] = \sum_{x \in S} \text{prob}(x) \times \text{surprise\_at}(x).$$

The question is how to compute the surprise. Note that the surprise depends only on the probability of an outcome. Note also that the surprise is decreasing with probability and that the surprise for independent events is additive. The surprise function also has the properties that there is no surprise at certain outcomes and infinite surprise at impossible outcomes. Finally, a small change in probability should result in a small change in the surprise. These properties lead us to the following axiomatization of a surprise function defined on probabilities. $h(\text{prob}(x))$ will then give the surprise_at$(x)$.

A **surprise function** h is a function $h : [0, 1] \rightarrow \Re$ that satisfies:

1. h(1) = 0 [No surprise at a certain event.]
2. h(0) = $\infty$ [Infinite surprise at an impossible event.]
3. $p_i > p_j \Rightarrow h(p_i) < h(p_j)$ [Surprise decreases with the probability of an event.]
4. $X, Y$ independent $\Rightarrow h(p(x_i \& y_j)) = h(p(x_i) * p(y_j)) = h(p(x_i)) + h(p(y_j))$ [Total surprise at independent events is the sum of the individual surprises.]
5. $h()$ is continuous [Small changes in probabilities lead to small changes in surprise.]

The *only* functions which satisfy the above five properties are functions of the form

$$h(p) = -k\, log_b(p) \text{ where } k, b > 0.$$

By convention we take $k = 1$ and $b = 2$ which yields $h(p) = -log_2(p)$. This surprise is measured in *bits* (because $b = 2$). The entropy or expected surprise is also a *lower bound* on *expected* number of yes/no questions needed to determine the answer.

Substituting this function back into the formula for entropy above, we find that if the random variable $X$ takes on the values $(x_1, x_2, ..., x_n)$ with probabilities $(p_1, p_2, ..., p_n)$, then the entropy is

$$H(X) = H(p_1, p_2, ..., p_n) = \sum_{i=1}^{n} p_i h(p_i) = -\sum_{i=1}^{n} p_i log_2(p_i).$$

*Some Definitions and Facts ...*

## Definitions:

**Entropy:** (information content)

$$H(X) \;=\; -\sum_{i \epsilon X} p(x_i) \log_2 [p(x_i)] \;=\; -E_x[log_2(p(x))]$$

**Joint Uncertainty:** (joint info, joint entropy)

$$H(X,Y) \;=\; -\sum_{i \epsilon X} \sum_{j \epsilon Y} p(x_i \& y_j) \log_2 [p(x_i \& y_j)] \;=\; -E_{x,y}[log_2(p(x,y))]$$

**Conditional Uncertainty:**

$$H(Y|X) \;=\; -\sum_{i \epsilon X} \sum_{j \epsilon Y} p(x_i \& y_j) \log_2 [p(y_j|x_i)] \;=\; -E_{x,y}[log_2(p(x|y))]$$

**Mutual Information:** (info about $X$ conveyed by $Y$)

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

**Relative Entropy:** (difference between two distributions)

$$D(p||q) \;=\; \sum_i p(x_i) \log_2 \left[\frac{p(x_i)}{q(x_i)}\right]$$

Relative Entropy obeys the property:

$$D(p||q) \;>=\; 0$$

with equality if and only if $p(x_i) \;=\; q(x_i) \quad \forall i$

# Facts:

$$H(p_1, p_2, \ldots, p_n) \;\leq\; \log_2(n) \quad (\;equality \Leftrightarrow p_i = \frac{1}{n}\;\forall i\;) \tag{1}$$

$$H(X,Y) \;\leq\; H(X) + H(Y) \quad (\;equality \Leftrightarrow X\&Y \text{ are independent}\;) \tag{2}$$

$$H(Y|X) \;\leq\; H(Y) \quad (\;equality \Leftrightarrow X\&Y \text{ are independent}\;) \tag{3}$$

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= \underbrace{H(Y)}_{entropy} + \underbrace{H(X|Y)}_{\substack{residual\\uncertainty}} \end{aligned} \tag{4}$$

$$H(X|Y) = 0 \;\Rightarrow\; H(Y) \geq H(X) \tag{5}$$

$$I(X|Y) \;\geq\; 0 \quad (\;equality \Leftrightarrow X\&Y \text{ are independent}\;) \tag{6}$$

$$I(X,Y) = I(X|Y) = I(Y|X) \tag{7}$$

Proof of fact (7): *( subtracting $H(Y|X) + H(X|Y)$ from each part of fact (4) )*

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) = H(Y) + H(X|Y) \\ H(X,Y) - H(Y|X) - H(X|Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= I(X|Y) = I(Y|X) = I(X,Y) \end{aligned}$$

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) && (8) \\ &= H(X,Y) - H(Y|X) - H(X|Y) && (9) \end{aligned}$$

Proof of fact (8): *( from the proof of fact (7) and fact (4) )*

$$\begin{aligned} I(X,Y) &= H(X,Y) - H(X|Y) - H(Y|X) \\ &= H(X,Y) - [H(X,Y) - H(Y)] - [H(X,Y) - H(X)] \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

$$\begin{aligned} D(p(x,y)||p(x)p(y)) &= \sum_x \sum_y p(x,y) \log_2 \left[\frac{p(x,y)}{p(x)p(y)}\right] && (10) \\ &= H(X) + H(Y) - H(X,Y) && (11) \\ &= I(X,Y) && (12) \end{aligned}$$

Thus, the *mutual information* between $X$ and $Y$ is precisely the relative entropy between the joint distribution of $X$ and $Y$ and the product distribution of $X$ and $Y$. Consequently, $D(p(x,y)||p(x)p(y))$ provides a measure of the excess coding cost incurred by assuming the independence of $X$ and $Y$, when in fact they are not independent.