Limits of Communication

Arnaldo Spalvieri

Dipartimento di Elettronica, Informazione e Bioingegneria

Politecnico di Milano, ITALY

Surprise, Entropy, Uncertainty as a Measure of Information

In 1928 Hartley proposed a measure of the *surprise* associated to a random variable. According to Hartley, a large surprise is associated to rare events. The surprise associated to the outcome x of the discrete random variable X is

$$H(x) = -\log_2 P(x) \geq 0.$$

One basic property of $H(x)$ is that the surprise of two independent events is the sum of the surprises of the individual events:

$$H(x, y) = -\log_2 P(x, y) = -\log_2 P(x)P(y) = -\log_2 P(x) - \log_2 P(y).$$

Taking the expectation of $H(x)$ one has the average surprise, or *uncertainty*, or *entropy* of the r.v. X:

$$H(X) = -\sum_{x \in \Omega_X} P(x) \log_2 P(x) \geq 0.$$

Example

Consider the binary space $\Omega_X = \{0, 1\}$.

$$H(X) = -P(1)\log_2 P(1) - (1 - P(1))\log_2(1 - P(1)) \leq 1.$$

$H(X)$ peaks at $P(1) = 0.5$: the uncertainty about the outcome of the binary r.v. is maximal when the two outcomes are equally likely. When $P(1) = 0$, or $P(1) = 1$ the entropy is $H(X) = 0$, as it should be since there is no uncertainty about the outcome. More generally, the entropy of a discrete random variable with K possible outcomes obeys to the inequality

$$H(X) \leq \log_2 K,$$

and the maximum is achieved when the outcomes are equally likely. In a lottery where one among $K = 2^n$ numbers can be extracted with uniform probability the entropy is just n, the number of bits that are used to encode the sample space of K numbers.

Residual Surprise or Conditional Entropy

Suppose that, ia a joint experiment $\{X, Y\}$, we observe the outcome of Y. The surprise that we have in knowing the outcome x of X after having observed y is

$$H(x|y) = -\log_2 P(x|y).$$

Of course, if there is complete dependence between X and Y, that is $x = y$, then there is zero surprise when, after having observed y, we observe x. On the opposite, if X and Y are independent, then $P(x|y) = P(x)$, and knowing y does not diminishes the surprise that we have when we have access to x. The average residual surprise about X after having observed y is

$$H(X|y) = -\sum_{x \in \Omega_X} P(x|y) \log_2 P(x|y).$$

Residual Surprise or Conditional Entropy

When there is complete dependence between X and Y, then $P(x|y)$ takes on only the values 0 and 1, therefore $H(X|y) = 0, \;\; \forall y \in \Omega_y$, meaning that there is zero residual surprise about X given the outcome y. In other words, the outcome of X is known with probability 1 when y is observed.

Averaging over the distribution of Y one gets

$$H(X|Y) = \sum_{y \in \Omega_Y} P(y) H(X|y) = -\sum_{x \in \Omega_X} \sum_{Y \in \Omega_y} P(y) P(x|y) \log_2 P(x|y),$$

which is a measure of the average uncertainty that remains about X after having observed Y. It can be seen that

$$H(X) \geq H(X|Y) \geq 0.$$

The upper bound is achieved when X and Y are independent, the lower bound when X is a deterministic and invertible function of Y.

Mutual Information

The information about X carried by the observation of Y is the difference between the uncertainty about X before observing Y and the residual uncertainty about X after having observed Y

$$I(X, Y) = H(X) - H(X|Y).$$

The following equality can be proved in a straightforward way

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

leading to the property

$$I(X, Y) = I(Y, X),$$

which explains the adjective *mutual* that is often used in front of *information*. It is easy to prove that

$$\min\{H(X), H(Y)\} \geq I(X, Y) \geq 0.$$

Mutual Information

Writing

$$H(X) = -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} P(x,y)\log_2 P(x),$$

one writes the mutual information as

$$I(X,Y) = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} P(x,y)\log_2\left(\frac{P(x,y)}{P(x)P(y)}\right).$$

The above formula is a measure of how much the probability in the numerator of the fraction inside the $\log$ diverges from the probability in the denominator.

In the extreme case of complete dependence, where $P(x,y) = P(x) = P(y)$, one has $I(X,Y) = H(X) = H(Y)$. At the opposite, when X and Y are independent, $P(x,y) = P(x)P(y)$ and $I(X,Y) = 0$.

Channel Capacity

Let X be the input to the channel and let Y be the output. From the information-theoretical standpoint, the channel is characterized by the conditional probability distribution $P(y|x)$.

The *unconstrained channel capacity* per channel use, or, in short, channel capacity, is

$$C = \max_{P(x)} I(X, Y), \quad \text{bits/channel use.}$$

For a fixed distribution of the source, $I(X, Y)$ is often referred to as *constrained* capacity.

The channel capacity theorem (Shannon, 1948) states that

given a channel with capacity C, it is possible to transmit through the channel $\beta \leq C$ bits per channel use with arbitrarily low probability of error.

Note that channel capacity, which involves a maximization over $P(x)$, is a function only of the channel transition probability $P(y|x)$, not of the source probability. In other words, the channel *is* $P(y|x)$.

Example: Capacity of the Binary Symmetric Channel (BSC)

Let $\Omega_x = \Omega_y = \{0, 1\}$, let $P(Y = 0|X = 1) = P(Y = 1|X = 0) = p$ be the channel transition probability, and let $P(X = 0) = q$, $P(X = 1) = 1 - q$. Then

$$H(Y|X) = -plog_2(p) - (1 - p)\log_2(1 - p)$$

is independent of q, and the capacity is found by maximizing $H(Y)$ versus q. Write

$$P(Y = 0) = q(1 - p) + (1 - q)p, \quad P(Y = 1) = (1 - q)(1 - p) + qp.$$

Since $H(Y)$ is maximum when $P(Y = 0) = P(Y = 1)$=0.5, we come to the conclusion that the capacity is achieved with $q = 0.5$, and that

$$C = 1 + plog_2(p) + (1 - p)\log_2(1 - p).$$

Entropy of a Continuous Random Variable

The entropy of the continuous random variable X, often called *differential* entropy is defined as

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log_2 f(x) dx.$$

The differential entropy of a continuous random variable with given variance σ_x^2 is upperbounded as

$$h(X) \leq \frac{1}{2} \log_2(2\pi e \sigma_x^2),$$

where equality holds when X is Gaussian.

Channels with Continuous Input and Continuous Output

A channel with continuous input and continuous output is a channel whose input is a continuous random variable X, and whose output is a continuous random variable Y. The conditional differential entropy that characterizes the channel is

$$h(Y|x) = -\int_{-\infty}^{\infty} f(y|x)\log_2(f(y|x))dy,$$

$$h(Y|X) = \int_{-\infty}^{\infty} f(x)h(Y|x)dx.$$

The mutual information is

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y,x)\log_2\left(\frac{f(y,x)}{f(x)f(y)}\right)dydx.$$

Capacity of the Additive Gaussian Channel

Consider the additive Gaussian noise channel

$$Y = X + N,$$

where X is continuous. The capacity of the channel with the constraint that the variance of X is fixed to σ_x^2 is readily obtained by considering that

$$h(Y|X) = h(N) = \frac{1}{2}\log_2(2\pi e \sigma_n^2),$$

and that $h(Y)$ is constrained to

$$h(Y) \leq \frac{1}{2}\log_2(2\pi e(\sigma_n^2 + \sigma_x^2)),$$

the maximum being achieved when Y is Gaussian, hence when X is Gaussian.

Capacity of the Additive Gaussian Channel

The capacity results

$$C = \frac{1}{2}\log_2(2\pi e(\sigma_n^2 + \sigma_x^2)) - \frac{1}{2}\log_2(2\pi e\sigma_n^2) = \frac{1}{2}\log_2(1 + \text{SNR}) \text{ bit/channel use.}$$

The result can be extended to a vector channel, where one channel use corresponds to a vector of N joint random variables (X_i, Y_i), $i = 1, 2, \cdots, N$. Assume that the pair (X_i, Y_i) is independent of the pair (X_j, Y_j), and that the channel is stationary. Invoking the basic property of the entropy of independent random variables one gets

$$C = \frac{N}{2}\log_2(1 + \text{SNR}) \text{ bit/vector channel use.}$$

Capacity of the AWGN (Additive White Gaussian Noise) Channel

Consider a channel of bandwidth B affected by additive Gaussian noise. The channel can be seen as a vector channel with one channel use per second, the vector having $N = 2B$ entries (recall the sampling theorem, B Hz=2B samples per second). Moreover, assume that the power spectral density of the noise is white, hence the samples of the Gaussian noise are independent of each other, leading to a vector channel with $N = 2B$ independent entries. The bit rate R_b that can be transferred through the channel with zero error probability is upper bounded as

$$R_b \leq B \log_2(1 + \text{SNR}) = C \quad \text{bit/s.}$$

Let $N_0/2$ be the power spectral density of the Gaussian noise. The power of the noise in the bandwidth B is $N_0 B$ and

$$R_b \leq B \log_2 \left(1 + \frac{P_x}{N_0 B}\right) \quad \text{bit/s.}$$

Spectral Efficiency

The *spectral efficiency*, that is hereafter denoted η, is the bit rate that is transferred through a channel of bandwidth 1 Hz in 1 second, therefore it is just

$$\eta = \frac{R_b}{B} \leq \log_2\left(1 + \frac{P_x}{N_0 B}\right).$$

η is a pure number, and it is expressed in bit/(Hz · s) or in bit/2D. By substituting $B = R_b/\eta$ inside the logarithm one gets

$$\eta \leq \log_2\left(1 + \frac{P_x \eta}{N_0 R_b}\right), \quad 2^\eta \leq 1 + \frac{P_x \eta}{N_0 R_b}, \quad \frac{P_x}{N_0 R_b} \geq \frac{2^\eta - 1}{\eta}.$$

Often the energy per bit $E_b = P_x/R_b$ is used in the last equation, leading to

$$\frac{E_b}{N_0} \geq \frac{2^\eta - 1}{\eta}.$$

Capacity of the AWGN Channel in the Power Constrained Region

Consider again the inequality

$$\frac{E_b}{N_0} \geq \frac{2^\eta - 1}{\eta}, \text{ or } \quad \frac{P_x}{R_b N_0} \geq \frac{2^{R_b/B} - 1}{R_b/B}.$$

Suppose that P_x is a cost, while B comes free. To exploit the infinite bandwidth we let $B \to \infty$, getting

$$\frac{P_x}{R_b N_0} \geq \lim_{B \to \infty} \frac{2^{R_b/B} - 1}{R_b/B} = \log_e 2 = -1.59 \text{ dB}.$$

This means that when $P_x/(R_b N_0)$ is below -1.59 dB it is impossible to transmit also with infinite bandwidth. Suppose that the signal power P_x is fixed (power constraint). Then what one can do is to renounce to some bit rate, until $P_x/(R_b N_0)$ goes above -1.59 dB. Conversely, if $P_x/(R_b N_0) > -1.59$ dB, then there is room to increase R_b. Often one uses the *energy per bit* E_b to expresses the ratio P_x/R_b. In this case

$$\frac{P_x}{R_b N_0} = \frac{E_b}{N_0}.$$

Capacity of the AWGN Channel in the Bandwidth Constrained Region

Consider capacity achieving transmission, that is

$$\eta = \log_2\left(1 + \frac{P_x}{N_0 B}\right).$$

Suppose that B and N_0 are fixed and that SNR is large, hence the $+1$ inside the log can be neglected. We want to know how much extra power we must add to increase the number of bits per complex channel use from η to $\eta + 1$. This is easily seen by noting that

$$2^{\eta} = \frac{P_x}{N_0 B}, \quad 2^{\eta+1} = \frac{2P_x}{N_0 B},$$

that is the law of 3 dB/bit. Besides channel capacity, also QAM follows the law of 3 dB/bit (not PSK), but with a gap in SNR from capacity that can be filled by channel coding. Exercise: compute the gap between 16-QAM and channel capacity, assuming that the symbol error rate of 10^{-6} is low enough to declare error free transmission. Repeat with 64-QAM and check the law.