Entropy

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x)$

$$p(x) = Pr\{X = x\}, \quad x \in \mathcal{X}$$

The *entropy* of the variable X is defined by

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

The logarithm can be in any base, but normally base 2 is used. The unit of the entropy is then called *bits*. If we use base b in the logarithm, we denote the entropy by $H_b(X)$. We can easily convert between entropies in different bases

$$H_b(X) = \log_b a \cdot H_a(X)$$

By convention $0 \log 0 = 0$, since $y \log y \to 0$ as $y \to 0$.

The entropy is a measure of the information content of a random variable.

Entropy, cont.

The entropy is always non-negative

$$H(X) \geq 0$$

Proof: Since $0 \leq p(x) \leq 1$, $-\log p(x) \geq 0$

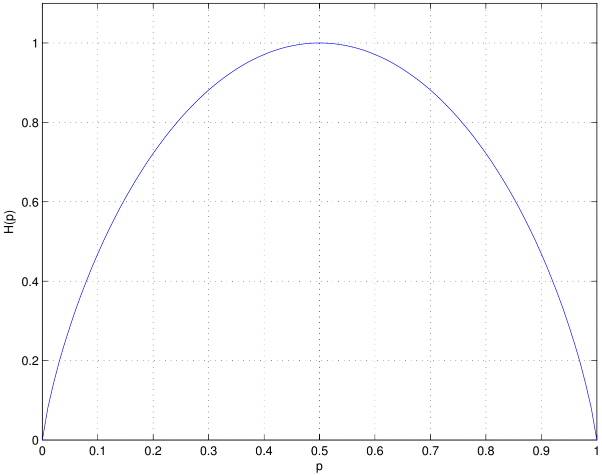
The entropy is always less than the logarithm of the alphabet size

$$H(X) \leq \log |\mathcal{X}|$$

with equality given a uniform distribution.

Proof: see later slide

Binary entropy



Entropy for a binary variable with symbol probabilities p and $1 - p$.
$H(p) = -p \cdot \log p - (1 - p) \cdot \log(1 - p)$

Joint and conditional entropy

The *joint entropy* of a pair of discrete random variables (X, Y) with joint probability mass function $p(x, y)$ is defined by

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

The *conditional entropy* of Y given X is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned}$$

Chain rule

The joint entropy can be written as the sum

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

Proof:

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Chain rule, cont.

We of course also have

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

The chain rule can easily be generalized to larger collections of random variables. Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then we have that

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

Relative entropy

The *relative entropy* or *Kullback-Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined by

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

We use the conventions $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

The relative entropy is always non-negative and zero if and only if $p = q$.

Note that the relative entropy is not a true metric, since it is not symmetric and does not satisfy the triangle inequality.

Mutual information

The *mutual information* between random variables X and Y with joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is defined as

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

The mutual information is a measure of the amount of information that one random variable contains about another random variable.

Mutual information, cont.

The mutual information can be rewritten as

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x|y)}{p(x)} \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y) \\ &= H(X) - H(X|Y) \end{aligned}$$

or symmetrically as

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

Thus X says as much about Y as Y says about X.

Mutual information, cont.

Note that

$$I(X;Y) = I(Y;X)$$

and

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

Using the chain rule, we also note that

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

Mutual information, cont.

The conditional mutual information of random variables X and Y given Z is defined by

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

Chain rule for mutual information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$

Jensen's inequality

A function f is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

The function is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.
A function f is (strictly) concave if $-f$ is (strictly) convex.

Jensen's inequality: If f is a convex function and X is a random variable then

$$Ef(X) \geq f(EX)$$

If f is strictly convex, then equality implies that $X = EX$ with probability 1, ie X is a constant.

Information inequality

Let $p(x), q(x), x \in \mathcal{X}$ be two probability mass functions. Then

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.
Proof: Let $A = \{x : p(x) > 0\}$. Then

$$\begin{aligned} -D(p||q) &= -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_{x\in A} p(x) \frac{q(x)}{p(x)} = \log \sum_{x \in A} q(x) \\ &\leq \log \sum_{x\in\mathcal{X}} q(x) = \log 1 = 0 \end{aligned}$$

Since $\log t$ is a strictly concave function of t, we have equality if and only if $q(x)/p(x)$ is constant everywhere, ie $p(x) = q(x)$

Mutual information, entropy

For any two random variables X and Y we have

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.
Proof: $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$ with equality if and only if $p(x,y) = p(x)p(y)$, ie X and Y are independent.

The entropy is bounded by the logarithm of the alphabet size

$$H(X) \leq \log |\mathcal{X}|$$

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over X and let $p(x)$ be the probability mass function for X. Then

$$D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X)$$

Since $D(p||u) \geq 0$ we get the inequality

Conditional entropy

Conditioning reduces entropy

$$H(X|Y) \leq H(X)$$

with equality of and only if X and Y are independent.

Proof: $0 \leq I(X;Y) = H(X) - H(X|Y)$

Knowing another random variable Y reduces (on average) the uncertainty of variable X.

Markov chains

Random variables X, Y, Z are said to *form a Markov chain in that order* (denoted $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X, ie if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

$X \to Y \to Z$ if and only if X and Z are conditionally independent given Y. Markovity gives

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

$X \to Y \to Z$ implies that $Z \to Y \to X$.
If $Z = f(Y)$ then $X \to Y \to Z$.

Data processing inequality

If $X \rightarrow Y \rightarrow Z$ then $I(X;Y) \geq I(X;Z)$
Proof: By the chain rule, we can expand mutual information in two ways

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

Since X and Z are conditionally independent given Y, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$ we get the inequality.
We have equality if and only if $I(X;Y|Z) = 0$, ie $X \rightarrow Z \rightarrow Y$) forms a Markov chain.

This means that no processing of Y, deterministic or random, can increase the amount of information that Y contains about X.

Law of large numbers

The law of large numbers states that for independent, identically distributed (i.i.d) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to the expected value EX for large n.

Definition: Given a sequence of random variables $X_1, X_2, \ldots$ we say that the sequence converges to a random variable X

1. *In probability* if for every $\epsilon > 0, Pr\{|X_n - X| > \epsilon\} \to 0$
2. *In mean square* if $E(X_n - X)^2 \to 0$
3. *With probability* (or *almost surely*) if $Pr\{\lim_{n\to\infty} X_n = X\} = 1$

Asymptotic equipartition property

If the random variables $X_1, X_2, \ldots$ are i.i.d $\sim p(x)$ then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability}$$

Proof: Functions of independent random variables are also independet random variables. Since X_i are i.i.d., so are $\log p(X_i)$. Thus, by the law of large numbers

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\sum_i \log p(X_i) \\ &\to -E\log p(X) \text{ in probability} \\ &= H(X) \end{aligned}$$

Typical sets

The *typical set* $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n)$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

The set $A_\epsilon^{(n)}$ has the following properties

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$ then $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$
2. $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for sufficiently large n
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for sufficiently large n

The typical set has probability near 1, all elements of the typical set are nearly equiprobable and the number of elements in the set is nearly 2^{nH}.

Typical sets, cont.

Property 1 follows directly from the definition of the typical set.

Property 2 follows from the AEP proof.

Property 3 follows from

$$\begin{aligned} 1 &= \sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x}) \geq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x}) \\ &\geq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)}|A_\epsilon^{(n)}| \end{aligned}$$

and thus $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

Typical sets, cont.

Finally, for sufficiently large n we have $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$. Thus property 4 follows from

$$1 - \epsilon < Pr\{A_\epsilon^{(n)}\} \leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|$$

and thus $|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}$.

Uses of entropy and mutual information

Lossless source coding

Assuming that we want to compress a sequence of i.i.d. random variables X_i. The entropy of the variables will limit the achievable rate R (in bits/symbol)

$$R \geq H(X_i)$$

Channel coding

Assuming that we want to transmit data over a discrete noisy channel, described by a conditional distribution $p(y|x)$. The capacity C of the channel (ie the maximum number of bits we can transmit over the channel per channel use and still be able to correct any errors) is given by maximizing the mutual information between the channel input and the channel output over all input distributions $p(x)$.

$$C = \max_{p(x)} I(X;Y)$$