Kraft's inequality

An instantaneous code (prefix code, tree code) with the codeword lengths $l_1, \ldots, l_N$ exists if and only if

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1$$

Proof: Suppose that we have a tree code. Let $l_{\max} = \max\{l_1, \ldots, l_N\}$. Expand the tree so that all branches have the depth $l_{\max}$. A codeword at depth l_i has $2^{l_{\max}-l_i}$ leaves underneath itself at depth $l_{\max}$. The sets of leaves under codewords are disjoint. The total number of leaves under codewords are less than or equal to $2^{l_{\max}}$. Thus we have

$$\sum_{i=1}^{N} 2^{l_{\max}-l_i} \leq 2^{l_{\max}} \Rightarrow$$

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1$$

Kraft's inequality, cont.

Conversely, given codeword lengths $l_1, \ldots, l_N$ that fulfill Kraft's inequality, we can always construct a tree code.

Start with a complete binary tree where all the leaves are at depth $l_{\max}$.

Assume, without loss of generality, that the codeword lengths are sorted in increasing order.

Choose a free node at depth l_1 for the first codeword and remove all its descendants. Do the same for l_2 and codeword 2, etc. until we have placed all codewords.

Kraft's inequality, cont.

Obviously we can place a codeword at depth l_1.

In order for the algorithm to work, there must in every step i be free leaves at the maximum depth $l_{\max}$. The number of remaining leaves is

$$2^{l_{\max}} - \sum_{j=1}^{i-1} 2^{l_{\max}-l_j} = 2^{l_{\max}}(1 - \sum_{j=1}^{i-1} 2^{-l_j}) > 2^{l_{\max}}(1 - \sum_{j=1}^{N} 2^{-l_j}) \geq 0$$

where we used the fact that Kraft's inequality is fulfilled.

This shows that there are free leaves in every step. Thus, we can construct a tree code code with the given codeword lengths.

Kraft-McMillan's inequality

Kraft's inequality can be shown to be fulfilled for all uniquely decodable codes, not just prefix codes. It is then called Kraft-McMillan's inequality: A uniquely decodable code with the codeword lengths $l_1, \ldots, l_N$ exists if and only if

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1$$

Consider $(\sum_{i=1}^{N} 2^{-l_i})^n$, where n is an arbitrary positive integer.

$$(\sum_{i=1}^{N} 2^{-l_i})^n = \sum_{i_1=1}^{N} \cdots \sum_{i_n=1}^{N} 2^{-(l_{i_1}+\ldots+l_{i_n})}$$

Kraft-McMillan's inequality, cont.

$l_{i_1} + \ldots + l_{i_n}$ is the length of n codewords from the code. The smallest value this exponent can take is n, which would happen if all code words had the length 1. The largest value the exponent can take is nl where l is the maximal codeword length. The summation can then be written as

$$(\sum_{i=1}^{N} 2^{-l_i})^n = \sum_{k=n}^{nl} A_k 2^{-k}$$

where A_k is the number of combinations of n codewords that have the combined length k. The number of possible binary sequences of length k is 2^k. Since the code is uniquely decodable, we must have

$$A_k \leq 2^k$$

in order to be able to decode.

Kraft-McMillan's inequality, cont.

We have

$$(\sum_{i=1}^{N} 2^{-l_i})^n \leq \sum_{k=n}^{nl} 2^k 2^{-k} = nl - n + 1$$

which gives us

$$\sum_{i=1}^{N} 2^{-l_i} \leq (n(l-1)+1)^{1/n}$$

This is true for all n, including when we let n tend to infinity, which finally gives us

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1$$

The converse of the inequality has already been proven, since we know that we can construct a prefix code with the given codeword lengths if they fulfill Kraft's inequality, and all prefix codes are uniquely decodable.

Instantaneous codes

One consequence of Kraft-McMillan's inequality is that it tells us that there is nothing to gain by using codes that are uniquely decodable but not instantaneous.

Given a uniquely decodable but not instantaneous code with codeword lengths $l_1, \ldots, l_N$, Kraft's inequality gives that we can always construct an instantaneous code with exactly the same codeword lengths. This new code will have the same rate as the old code but it will be easier to decode.

Performance measure

We measure how good a code is with the *mean data rate* R (more often just *data rate* or *rate*).

$$R = \frac{\text{average number of bits per codeword}}{\text{average number of symbols per code word}} \quad \text{[bits/symbol]}$$

Since we're doing data compression, we want the rate to be as low as possible.
If we initially assume that we have a memoryless source X_j and code one symbol at a time using a tree code, then R is given by

$$R = \bar{l} = \sum_{i=1}^{L} p_i \cdot l_i \quad \text{[bits/symbol]}$$

where L is the alphabet size and p_i the probability of symbol i.
$\bar{l}$ is the *mean codeword length* [bits/codeword].

Theoretical lower bound

Given that we have a memoryless source X_j and that we code one symbol at a time with a prefix code. Then the mean codeword length $\bar{l}$ (which is equal to the rate) is bounded by

$$\bar{l} \geq -\sum_{i=1}^{L} p_i \cdot \log_2 p_i = H(X_j)$$

$H(X_j)$ is called the *entropy* of the source.

Theoretical lower bound, cont.

Consider the difference between the entropy and the mean codeword length

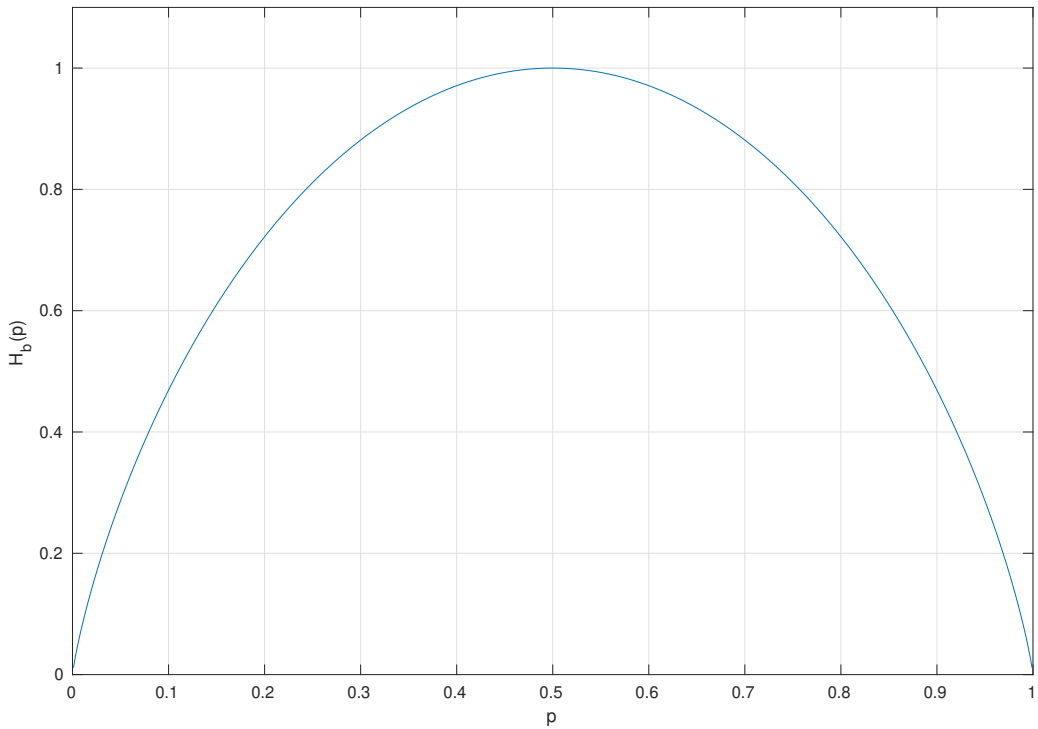
$$
\begin{aligned}
H(X_j) - \bar{l} &= -\sum_{i=1}^{L} p_i \cdot \log p_i - \sum_{i=1}^{L} p_i \cdot l_i = \sum_{i=1}^{L} p_i \cdot (\log \frac{1}{p_i} - l_i) \\
&= \sum_{i=1}^{L} p_i \cdot (\log \frac{1}{p_i} - \log 2^{l_i}) = \sum_{i=1}^{L} p_i \cdot \log \frac{2^{-l_i}}{p_i} \\
&\leq \frac{1}{\ln 2} \sum_{i=1}^{L} p_i \cdot (\frac{2^{-l_i}}{p_i} - 1) = \frac{1}{\ln 2} (\sum_{i=1}^{L} 2^{-l_i} - \sum_{i=1}^{L} p_i) \\
&\leq \frac{1}{\ln 2} (1 - 1) = 0
\end{aligned}
$$

where we used the fact that $\ln x \leq x - 1$ and Kraft's inequality.

Simplest example

Let X be a binary random variable with alphabet $\mathcal{A} = \{a, b\}$ and probabilities $p(a) = p$ and $p(b) = 1 - p$. The entropy of X is

$$H(X) = -p \cdot \log p - (1 - p) \cdot \log(1 - p) = H_b(p)$$



$H_b(p)$ is referred to as the *binary entropy function*.

Shannon's information measure

We want a measure I of information that is connected to the probabilities of events.
Some desired properties:

- The information $I(A)$ of an event A should only depend of the probability $P(A)$ of the event.
- The lower the probability of the event, the larger the information should be.
- If the probability of an event is 1, the information should be 0.
- Information should be a continuous function of the probability.
- If the independent events A and B happen, the information should be the sum of the informations $I(A) + I(B)$

This gives that information should be a logarithmic measure.

Information Theory

The *information* $I(A;B)$ that is given about an event A, when event B happens is defined as

$$I(A;B) \triangleq \log_b \frac{P(A|B)}{P(A)}$$

where we assume that $P(A) \neq 0$ and $P(B) \neq 0$. In the future we assume, unless otherwise specified, that $b = 2$. The unit of information is then called *bit*. (If $b = e$ the unit is called *nat*.)
$I(A;B)$ is symmetric in A and B:

$$I(A;B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(AB)}{P(A)P(B)} =$$

$$= \log \frac{P(B|A)}{P(B)} = I(B;A)$$

Therefore the information is also called *mutual information*.

Information Theory, cont.

We further have that

$$-\infty \leq I(A;B) \leq -\log P(A)$$

with "equality" to the left if $P(A|B) = 0$ and equality to the right if $P(A|B) = 1$.
$I(A;B) = 0$ means that the events A and B are independent.
$-\log P(A)$ is the amount of information that needs to be given in order for us to determine that event A has happened.

$$I(A;A) = \log \frac{P(A|A)}{P(A)} = -\log P(A)$$

We define the *self information* of the event A as

$$I(A) \stackrel{\triangle}{=} -\log P(A)$$

Information Theory, cont.

If we apply the definitions on the events $\{X = x\}$ and $\{Y = y\}$ we get

$$I(X = x) = -\log p_X(x)$$

and

$$I(X = x; Y = y) = \log \frac{p_{X|Y}(x|y)}{p_X(x)}$$

These are real functions of the random variables X and the random variable (X, Y), so the mean values are well defined

$$H(X) \stackrel{\triangle}{=} E\{I(X = x)\} = -\sum_{i=1}^{L} p_X(x_i) \log p_X(x_i)$$

is called the *entropy* (or *uncertainty*) of the random variable X.

Information Theory, cont.

$$
\begin{aligned}
I(X;Y) &\stackrel{\triangle}{=} E\{I(X=x;Y=y)\} = \\
&= \sum_{i=1}^{L}\sum_{j=1}^{M} p_{XY}(x_i,y_j)\log\frac{p_{X|Y}(x_i|y_j)}{p_X(x_i)}
\end{aligned}
$$

is called the *mutual information* between the random variables X and Y.

Information Theory, cont.

If (X, Y) is viewed as *one* random variable we get

$$H(X, Y) = -\sum_{i=1}^{L}\sum_{j=1}^{M} p_{XY}(x_i, y_j) \log p_{XY}(x_i, y_j)$$

This is called the *joint entropy* of X and Y.
It then follows that the mutual information can be written as

$$\begin{aligned} I(X;Y) &= E\{\log \frac{p_{X|Y}}{p_X}\} = E\{\log \frac{p_{XY}}{p_X p_Y}\} \\ &= E\{\log p_{XY} - \log p_X - \log p_Y\} \\ &= E\{\log p_{XY}\} - E\{\log p_X\} - E\{\log p_Y\} \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

Information Theory, cont.

Useful inequality

$$\log r \leq (r-1)\log e$$

with equality if and only if $r = 1$.
This can also be written

$$\ln r \leq r - 1$$

If X takes values in $\{x_1, x_2, \ldots, x_L\}$ we have that

$$0 \leq H(X) \leq \log L$$

with equality to the left if and only if there is an i such that $p_X(x_i) = 1$ and with equality to the right if and only if $p_X(x_i) = 1/L$ for all $i = 1, \ldots, L$.

Information Theory, cont.

Proof left inequality:

$$-p_X(x_i) \cdot \log p_X(x_i) \begin{cases} = 0, & p_X(x_i) = 0 \\ > 0, & 0 < p_X(x_i) < 1 \\ = 0, & p_X(x_i) = 1 \end{cases}$$

Thus we have that $H(X) \geq 0$ with equality if and only if $p_X(x_i)$ is either 0 or 1 for each i, but this means that $p_X(x_i) = 1$ for exactly one i.

Information Theory, cont.

Proof right inequality:

$$\begin{aligned} H(X) - \log L &= -\sum_{i=1}^{L} p_X(x_i) \log p_X(x_i) - \log L \\ &= \sum_{i=1}^{L} p_X(x_i) \log \frac{1}{L \cdot p_X(x_i)} \\ &\leq \sum_{i=1}^{L} p_X(x_i) \left(\frac{1}{L \cdot p_X(x_i)} - 1\right) \log e \\ &= \left(\sum_{i=1}^{L} \frac{1}{L} - \sum_{i=1}^{L} p_X(x_i)\right) \log e \\ &= (1-1) \log e = 0 \end{aligned}$$

with equality if and only if $p_X(x_i) = \frac{1}{L}$ for all $i = 1, \ldots, L$

Information Theory, cont.

The *conditional entropy* of X given the event $Y = y_j$ is

$$H(X|Y = y_j) \stackrel{\triangle}{=} -\sum_{i=1}^{L} p_{X|Y}(x_i|y_j) \log p_{X|Y}(x_i|y_j)$$

We have that

$$0 \leq H(X|Y = y_j) \leq \log L$$

The conditional entropy of X given Y is defined as

$$\begin{aligned} H(X|Y) &\stackrel{\triangle}{=} E\{-\log p_{X|Y}\} = \\ &= -\sum_{i=1}^{L}\sum_{j=1}^{M} p_{XY}(x_i, y_j) \log p_{X|Y}(x_i|y_j) \end{aligned}$$

We have that

$$0 \leq H(X|Y) \leq \log L$$

Information Theory, cont.

We also have that

$$H(X|Y) = -\sum_{i=1}^{L}\sum_{j=1}^{M} p_{XY}(x_i, y_j) \log p_{X|Y}(x_i|y_j)$$

$$= -\sum_{i=1}^{L}\sum_{j=1}^{M} p_Y(y_j) p_{X|Y}(x_i|y_j) \log p_{X|Y}(x_i|y_j)$$

$$= -\sum_{j=1}^{M} p_Y(y_j) \sum_{i=1}^{L} p_{X|Y}(x_i|y_j) \log p_{X|Y}(x_i|y_j)$$

$$= \sum_{j=1}^{M} p_Y(y_j) H(X|Y = y_j)$$

Information Theory, cont.

We have

$$p_{X_1 X_2 \ldots X_N} = p_{X_1} \cdot p_{X_2|X_1} \cdots p_{X_N|X_1 \ldots X_{N-1}}$$

which leads to the *chain rule*

$$H(X_1 X_2 \ldots X_N) =$$

$$H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X_1 \ldots X_{N-1})$$

We also have that

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

Information Theory, cont.

Other interesting inequalities

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

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

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

with equality if and only if X and Y are independent.
If $f(X)$ is a function of X, we have

$$H(f(X)) \leq H(X)$$

$$H(f(X)|X) = 0$$

$$H(X, f(X)) = H(X)$$

Entropy for sources

The *entropy*, or *entropy rate* of a stationary random source X_n is defined as

$$\lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n) =$$
$$\lim_{n\to\infty} \frac{1}{n} (H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1 \dots X_{n-1})) =$$
$$\lim_{n\to\infty} H(X_n|X_1 \dots X_{n-1})$$

For a memoryless source, the entropy rate is equal to $H(X_n)$.

Entropy for Markov sources

The entropy rate of a stationary Markov source X_n of order k is given by

$$H(X_n|X_{n-1}\dots X_{n-k})$$

The entropy rate of the state sequence S_n is the same as the entropy rate of the source

$$\begin{aligned} H(S_n|S_{n-1}S_{n-2}\dots) &= H(S_n|S_{n-1}) = \\ H(X_n\dots X_{n-k+1}|X_{n-1}\dots X_{n-k}) &= H(X_n|X_{n-1}\dots X_{n-k}) \end{aligned}$$

and thus the entropy rate can also be calculated by

$$H(S_n|S_{n-1}) = \sum_{j=1}^{L^k} w_j \cdot H(S_n|S_{n-1} = s_j)$$

ie a weighted average of the entropies of the outgoing probabilities of each state.