MA3466 – Information Theory

David Whyte dawhyte@tcd.ie

Updated April 3, 2012

Contents

1 Probability & statistics

1.1 Random variables

When we say X is a random variable, we consider a set of outcomes (or "alphabet") $\mathscr{X} = \{x_1, x_2, \ldots, x_n\}$ where n is finite. It is not much more difficult to deal with the infinite discrete case – but the continuous case is more awkward in information theory. In this course we'll deal with the discrete case only.

For the set of outcomes $\mathscr{X}$, we have a map:

$$\begin{aligned} p_x : \mathscr{X} &\to \mathbb{R}_+ \\ x_i &\mapsto p_X(x_i) \end{aligned}$$

with

$$\sum_{x_i \in \mathscr{X}} p_X(x_i) = 1.$$

$p_X(x)$ can also be denoted $\Pr(x)$.

For example, if X is the roll of a die, then:

$$\mathscr{X} = \{1,2,3,4,5,6\}$$

$$p_X(1) = \ldots = p_X(6) = \tfrac{1}{6}.$$

If X is the next letter in a text,

$$\mathscr{X} = \{\mathsf{a}, \ldots, \mathsf{z}\}.$$

Here X is not random, but it is unpredictable!

The estimator for the probability is:

$$p_X(x_i) = \lim_{N \to \infty} \frac{\#\text{ times } x_i \text{ happens}}{N}$$

Generally:

$$\Pr(E) = \sum_{x \in \mathscr{X} \mid E \text{ holds}} p_X(x)$$

For example,

$$\Pr(\text{3 or less}) = p_X(1) + p_X(2) + p_X(3) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}$$

1.2 Expected value

Say $g : \mathscr{X} \to \mathbb{R}$ (or, more generally, any field). The expected value of g is:

$$E_X g = \langle g(X) \rangle_{p_X} = \sum_{x \in \mathscr{X}} g(x) p_X(x)$$

For example, if the distribution of heights in cm were:

Height	199	200	201
$\boldsymbol{p}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

To find the expected value of height, g is the trivial map and:

$$\langle X \rangle = \tfrac{1}{4} \cdot 199 + \tfrac{1}{2} \cdot 200 + \tfrac{1}{4} \cdot 201 = 200.$$

Or to consider a farmyard example,

Item	cow	sheep	turnip
$\boldsymbol{p}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Using the trivial map makes no sense since $\mathscr{X}$ has no additive structure. However if we define a map $€ : \mathscr{X} \to \mathbb{R}$, with €(cow) = 700, €(sheep) = 100 and €(turnip) = 0.2, for example, then:

$$\begin{aligned} \langle €(X) \rangle &= \tfrac{1}{2} \cdot 700 + \tfrac{1}{4} \cdot 100 + \tfrac{1}{4} \cdot 0.2 \\ &= 375.05 \end{aligned}$$

In statistics, we have an estimator for $\langle g(x) \rangle$:

$$\langle g(x) \rangle = \lim_{N \to \infty} \frac{1}{N} \sum_{\text{trials}} g(x)$$

If $\mathscr{X}$ does have a field structure, e.g. $\mathscr{X} \subset \mathbb{R}$, $\langle X \rangle$ is called the mean:

$$\langle X \rangle = \sum_{x \in \mathscr{X}} x p_X(x) = \overline{X}.$$

1.3 Moments

The n^{th} central moment is defined as:

$$\mu_n = \langle (X - \overline{X})^n \rangle.$$

Let's calculate μ_1 and μ_2:

$$\begin{aligned} \mu_1 &= \langle X - \overline{X} \rangle \\ &= \sum_{x \in \mathscr{X}} p_X(x) x - \sum_{x \in \mathscr{X}} p_X(x) \overline{X} \\ &= \overline{X} - \overline{X} \underbrace{\sum_{x \in \mathscr{X}} p_X(x)}_{1} \\ &= 0. \end{aligned}$$

$$\begin{aligned} \mu_2 &= \langle (X - \overline{X})^2 \rangle \\ &= \langle X^2 \rangle - 2\overline{X} \langle X \rangle + \underbrace{\langle \overline{X}^2 \rangle}_{\overline{X}^2 \langle 1 \rangle} \\ &= \langle X^2 \rangle - \overline{X}^2. \end{aligned}$$

$\mu_2 = \sigma^2$, the variance. σ is the standard deviation.

$\gamma_1 = \dfrac{\mu_3}{\sigma^3}$ is the skewness, and $\gamma_2 = \dfrac{\mu_4}{\sigma^4} - 3$ is the kurtosis.

In passing, the issue of estimators is non-trivial:

$$\overline{X} = \frac{1}{N} \sum_{N \text{ trials}} x$$

$$\sigma^2 = \frac{1}{N-1} \sum_{N \text{ trials}} (x - \overline{X})^2$$

Where the $N-1$ comes from is very subtle indeed! Notice that as $N \to \infty$ it becomes unimportant.

1.4 Multiple random variables

Say you have two random variables X and Y, with sets of outcomes $\mathscr{X}$ and $\mathscr{Y}$. We can consider $\mathscr{X} \times \mathscr{Y}$ and it is often possible to define the random variable (X, Y) with probabilities $p_{X,Y}(x, y)$.

For example, if X is the first letter of a random word and Y is the last letter of the same word,

$$p_X(\mathsf{e}) \approx 0.0186$$
$$p_Y(\mathsf{e}) \approx 0.1916$$

(for a certain corpus).

The probability that both the first and last letters are e is:

$$p_{(X,Y)}(\mathsf{e}, \mathsf{e}) \approx 0.0028$$

Note:

$$p_X(\mathsf{e}) p_Y(\mathsf{e}) = 0.0035 > p_{(X,Y)}(\mathsf{e}).$$

So X and Y are not independent – a word starting with e is less likely to end in e than a randomly selected word, and vice versa.

$p_{(X,Y)}(x, y)$ is called the *joint distribution*, and $p_X(x)$ and $p_Y(y)$ are called *marginal distributions*. $p_X(x)$ can be expressed as:

$$p_X(x) = \sum_{y \in \mathscr{Y}} p_{X,Y}(x, y).$$

There are also *conditional distributions* – for example, $p_{X|Y}(x|y)$ is the probability that $X = x$ given that $Y = y$.

1.5 Bayes' rule

Bayes' rule relates joint, conditional and marginal probabilities. One statement of the rule is:

$$\underbrace{p_{X,Y}(x,y)}_{\text{joint}} = \underbrace{p_{X|Y}(x|y)}_{\text{conditional}}\,\underbrace{p_Y(y)}_{\text{marginal}},$$

i.e. "The probability of getting $X = x$ and $Y = y$ is the probability of getting $Y = y$ multiplied by the probability of getting $X = x$ given that $Y = y$."

Switching $X \leftrightarrow Y$ gives: $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$, so we can combine these to give a more common statement of Bayes' rule.

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

Omitting subscripts (as above) is a common but potentially confusing abuse of notation.

Using the first/last letter example again, $p_{X,Y}(\mathsf{e},\mathsf{e}) = 0.0028$ and $p_X(\mathsf{e}) = 0.0186$. So:

$$p_{Y|X}(\mathsf{e},\mathsf{e}) = \frac{p_{X,Y}(\mathsf{e},\mathsf{e})}{p_X(\mathsf{e})} = 0.1522.$$

Compare this to $p_Y(\mathsf{e}) = 0.1916$.

For another example, consider $\mathscr{Y} = \{a,b\}$ and $\mathscr{X} = \{1,2,3\}$, with probabilities:

	a	***b***
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{12}$	$\frac{1}{6}$
3	$\frac{1}{6}$	$\frac{1}{4}$

We can calculate the marginal distributions for X and Y by simple addition:

X	
1	$\frac{1}{3}$
2	$\frac{1}{4}$
3	$\frac{5}{12}$

Y	
a	$\frac{5}{12}$
b	$\frac{7}{12}$

If we wanted to find $p_{X|Y}(1|b)$, we use Bayes' rule:

$$p_{X|Y}(1|b) = \frac{p_{X,Y}(1,b)}{p_Y(b)} = \frac{\frac{1}{6}}{\frac{7}{12}} = \frac{2}{7}.$$

1.6 Monty Hall Problem

If C is the position of the car, and R is the door opened by Monty, we can write:

$$\mathscr{C} = \{1,2,3\}, \quad \mathscr{R} = \{2,3\}$$

assuming that the contestant always chooses door 1 for simplicity.

Let's say $R = 2$ and we want $P_{C|R}(c|2)$. Use Bayes' rule:

$$p_{C|R}(c|r) = \frac{p_{R|C}(r|c)p_C(c)}{p_R(r)}$$

We know that $p_R(2) = \frac{1}{2}$ and $p_C(1) = \frac{1}{3}$. We also know the conditional probabilities $p_{R|C}(2|3) = 1$ and $p_{R|C}(2|1) = \frac{1}{2}$. So:

$$p_{C|R}(1|2) = \frac{p_{R|C}(2|1)p_C(1)}{p_R(2)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{2},$$
$$p_{C|R}(3|2) = \frac{p_{R|C}(2|3)p_C(3)}{p_R(2)} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

So the contestant is better off switching!

Also, $p_{C|R}(2|2) = 0$. Naturally.

2 Quantities in information theory

2.1 Entropy

Shannon's entropy is the central quantity in information theory. It is defined as:

$$\begin{aligned} H(X) &= -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x) \\ &= -\langle \log_2 p_X(x)\rangle_p. \end{aligned}$$

Note: "log" without a subscript is in this course taken to mean $\log_2$.

This definition can be justified somewhat – no matter what type of thing x is, $p_X(x)$ is always a number and so $\langle p_X(x)\rangle$ is well-defined. In addition, the log simplifies the multiplicative structure of probabilities into an additive one.

For example, for this uniform distribution:

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

the entropy is easily calculated:

$$\begin{aligned} H(x) &= -\tfrac{1}{8}\log\tfrac{1}{8} - \tfrac{1}{8}\log\tfrac{1}{8} - \ldots - -\tfrac{1}{8}\log\tfrac{1}{8} \\ &= -\log\tfrac{1}{8} \\ &= 3 \text{ bits}. \end{aligned}$$

(Information entropy is measured in bits).

For a different distribution:

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

the entropy is different:

$$\begin{aligned} H(x) &= -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{8}\log\tfrac{1}{8} - \tfrac{1}{16}\log\tfrac{1}{16} - 4\cdot\tfrac{1}{64}\log\tfrac{1}{64} \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{8} + \tfrac{4}{16} + \tfrac{24}{64} \\ &= 2 \text{ bits.} \end{aligned}$$

Lower entropy means that we learn less from hearing the 'answer' for this distribution – half the time, the answer is **a** so we are less surprised when it comes up. In the uniform distribution, all answers are equally surprising.

An efficient encoding scheme for the second distribution is:

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
0	10	110	1110	111100	111101	111110	111111

No 'comma' is needed – a string of bits can be uniquely resolved into characters.

The average length of one letter is:

$$\langle L\rangle = \tfrac{1}{2}\cdot 1 + \tfrac{1}{4}\cdot 2 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{16}\cdot 4 + 4\cdot\tfrac{1}{64}\cdot 6 = 2 \text{ bits.}$$

A naïve code such as:

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
000	001	010	011	100	101	110	111

which again needs no comma, would have $\langle L\rangle = 3$. This code is however optimal for the uniform distribution.

The average code length is closely related to the entropy, but they are not always equal as in these examples.

Lemma. $H(x) \geqslant 0$.

Proof. $-\log p_X(x) \geqslant 0$, since $0 \leqslant p_X(x) \leqslant 1$. □

Lemma. $H_b(x) = \log_b a H_a(x)$.

Proof.

$$\log_a B = \frac{\log_c B}{\log_c A}$$

$$\begin{aligned} \implies \log_b a H_a(x) &= -\sum_x p(x)\log_a b \log_a x \\ &= -\sum_x p(x)\log_b x \\ &= H_b(x). \end{aligned}$$

□

2.2 Joint entropy

If $p_{X,Y}(x,y)$ is a joint distribution, the *joint entropy* is defined as:

$$H(X,Y) = -\sum p_{X,Y}(x,y)\log p_{X,Y}(x,y),$$

i.e. the normal definition applied to a joint distribution.

2.3 Conditional entropy

The entropy of a conditional distribution $p_{X|Y}(x|y)$ is:

$$H(X|Y=y) = -\sum_{x\in\mathcal{X}} p_{X|Y}(x|y)\log p_{X|Y}(x|y).$$

The *conditional entropy* is the average of the entropies of the conditional distributions:

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

since $p_Y(y)p_{X|Y}(x|y) = p_{X,Y}(x,y)$ by Bayes' rule.

In other words,

$$H(X|Y) = -\langle \log p_{X|Y}(x|y)\rangle_{p_{X,Y}}$$

2.4 Chain rule for entropy

Theorem.

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

Proof.

$$\begin{aligned} H(X,Y) &= -\sum p_{X,Y}(x,y)\log p_{X,Y}(x,y) \\ &= -\sum p_{X,Y}\log\left[p_Y(y)p_{X|Y}(x|y)\right] \\ &= -\sum p_{X,Y}\log p_Y(y) \underbrace{-\sum p_{X,Y}\log p_{X|Y}(x|y)}_{H(X|Y)} \end{aligned}$$

As for the first term,

$$\begin{aligned} \sum_{x,y} p_{X,Y}(x,y)\log p_Y(y) &= \sum_y \log p_Y(y)\underbrace{\sum_x p_{X,Y}(x,y)}_{p_Y(y)} \\ &= \sum_y p_Y(y)\log p_Y(y) \\ &= H(Y). \end{aligned}$$

□

In words, the chain rule states: "The information in X and Y is equal to the information in Y plus the further information in X given that you know Y."

A corollary is that:

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

2.5 Kullback–Leibler divergence

Say $\mathscr{X} = \mathscr{Y}$, $p(x) \equiv p_X(x)$, $q(x) \equiv p_Y(x)$. The *Kullback–Leibler divergence* (or *KL divergence*, sometimes *relative entropy*) between p and q is:

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

In other words,

$$D(p\|q) = \left\langle \log \frac{p(x)}{q(x)} \right\rangle_p.$$

The important properties of the KL divergence are:

$$\begin{aligned} D(p\|q) &\geqslant 0 \quad \forall p,q \\ D(p\|q) &= 0 \quad \iff p(x) = q(x) \forall x \in \mathscr{X}. \end{aligned}$$

The KL distribution is somewhat like a 'distance' between two probability distributions. Note that it is *not* a metric – it is not symmetric!

In many applications of the KL divergence, you will have a "real" distribution and a "model" distribution, and want to minimise the divergence between them. It is commonly used in AI, where it is often as practical as the standard L^2 distance:

$$D_2 = \sqrt{\sum \big(p(x) - q(x)\big)^2}$$

One practical problem with the KL divergence is that if $p(x) \neq 0$ and $q(x) = 0$ for some $x \in \mathscr{X}$, then $D(p\|q)$ is undefined.

2.6 Mutual information

The *mutual information* between X and Y is:

$$I(X;Y) = D\big(p_{X,Y}(x,y)\|p_X(x)p_Y(y)\big),$$

or "how far the distributions are from being independent". Or using the definition of D,

$$I(X;Y) = \sum p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}.$$

2.7 Relationship between all this stuff

$$\begin{aligned} I(X;Y) &= \sum p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum p(x,y)\log\frac{p(x)p(y|x)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\log p(y|x) \underbrace{-\sum_{x,y} p(x,y)\log p(y)}_{=-\sum_y p(y)\log p(y)} \\ &= -H(Y|X) + H(Y). \end{aligned}$$

In other words, $H(Y) = H(Y|X) + I(X;Y)$, "the information in Y equals the information left in Y when you know X plus the information knowing X gives you about Y."

Another form is $I(X;Y) = H(X) + H(Y) - H(X,Y)$, so it is somehow a measure of the *overlap* of X and Y.

2.8 Chain rule for entropy

Theorem.

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

For example,

$$H(X_1, X_2, X_3) = H(X_1|X_2, X_3) + H(X_2|X_3) + H(X_3).$$

Proof.

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$\begin{aligned} \implies H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_3|X_2, X_1) + H(X_2|X_1) \end{aligned}$$

and so on. □

2.9 Chain rule for mutual information

We define *conditional mutual information*:

$$I(X;Y|Z) = \sum p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)},$$

the mutual information of the conditional probabilitied averaged over the condition.

By Bayes' rule we can write:

$$\begin{aligned} I(X;Y|Z) &= \sum p(z)\sum p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \sum p(z) I(X;Y|Z=z) \end{aligned}$$

Theorem.

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

Proof.

$$\begin{aligned} I(X_1, \ldots, X_n; Y) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n | Y) \\ &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y) \\ &= \sum_{i=1}^{n} \left[H(X_i | X_{i-1}, \ldots, X_1) - H(X_i | X_{i-1}, \ldots, X_1, Y) \right] \\ &= \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1) \end{aligned}$$

□

2.10 Chain rule for KL divergence

We define *conditional KL divergence*:

$$D\big(p(y|x) \| q(y|x)\big) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

This is just the average of the KL divergence of the conditional distributions. By Bayes' rule it can also be written:

$$\sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)}$$

Theorem.

$$D\big(p(x,y) \| q(x,y)\big) = D\big(p(x) \| q(x)\big) + D\big(p(y|x) \| q(y|x)\big)$$

Proof.

$$\begin{aligned} D\big(p(x,y) \| q(x,y)\big) &= \sum p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum p(x,y) \log \frac{p(x)}{q(x)} + \sum p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum p(x,y) \log \frac{p(x)}{q(x)} + \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= D\big(p(x) \| q(x)\big) + D\big(p(y|x) \| q(y|x)\big). \end{aligned}$$

□

3 Jensen's inequality etc.

3.1 Cup-like & cap-like functions

Eschewing the terms "convex" and "concave", we say that a function f is *cup-like* if for all $a < b$, $\lambda \in (0,1)$:

$$f\big((1-\lambda)a + \lambda b\big) \leqslant (1-\lambda) f(a) + \lambda f(b),$$

i.e. the function is below the straight line, as in Fig. 1.

We say f is strictly cup-like if the inequality is strict, and f is *cap-like* if $-f$ is cup-like.

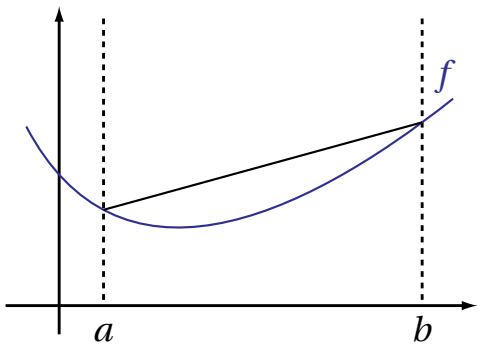


Figure 1: A cup-like function

3.2 Jensen's inequality

Theorem. *If f is cup-like on X,*

$$\langle f(X)\rangle \geqslant f(\langle X\rangle).$$

Furthermore if f is strictly cup-like,

$$\langle f(X)\rangle = f(\langle X\rangle) \iff X = \langle X\rangle,$$

i.e. X is constant.

Proof. Take the case $\mathscr{X} = \{x_1, x_2\}$ in which case the inequality follows directly from the definition of cup-like, replacing $\lambda \to p_1$:

$$p_1 f(x_1) + p_2 f(x_2) \geqslant f(p_1 x_1 + p_2 x_2),$$

Now assume that the inequality holds for $\mathscr{X} = \{x_1, \ldots, x_{n-1}\}$. Assuming $p_n \neq 1$, let

$$p'_i = \frac{p_i}{1-p_n},$$

so that $(p'_1,\dots,p'_n)$ is a probability distribution on $\mathscr{X}' = \{x_1,\dots,x_{n-1}\}$. So:

$$\begin{aligned}
\sum_{i=1}^{n} p_i f(x_i) &= p_n f(x_n) + (1-p_n)\sum_{i=1}^{n-1} p'_i f(x_i) \\
&\geqslant p_n f(x_n) + (1-p_n) f\left(\sum p'_i x_i\right) && \text{(by induction assumption)} \\
&\geqslant f\left(p_n x_n + (1-p_n)\sum p'_i x_i\right) && \text{(by definition of cup-like)} \\
&= f\left(\sum p_i x_i\right)
\end{aligned}$$

i.e. $\langle f(X)\rangle \geqslant f(\langle X\rangle)$.

If $f(x)$ is strictly cup-like, the 2-point case reads:

$$f(p_1x_1 + p_2x_2) = p_1 f(x_1) + p_2 f(x_2) \iff p_1 = 1 \text{ or } p_2 = 1,$$

and proceed as above. □

3.3 Information inequality

This is also called *Gibbs' inequality*.

Theorem. *Let $p(x)$ and $q(x)$ be probability distributions on $\mathscr{X}$. Then,*

$$D(p\|q) \geqslant 0,$$

with $D(p\|q) = 0 \iff p(x) = q(x) \forall\, x \in \mathscr{X}$.

Proof. Let $\mathscr{A}$ be the support of p: $\mathscr{A} = \{x \in \mathscr{X} : p(x) \neq 0\}$. We can write:

$$-D(p\|q) = -\sum_{x\in\mathscr{A}} p(x)\log\frac{p(x)}{q(x)}$$

Now define a new random variable Y:

$$y_i = \frac{q(x_i)}{p(x_i)}, \text{ with } p_Y(y_i) = p_X(x_i).$$

We want to apply Jensen's inequality to this new variable.

$$\begin{aligned}
-D(p\|q) &= -\sum_{x\in\mathscr{A}} p(x)\log\frac{p(x)}{q(x)} \\
&= \langle \log Y\rangle \\
&\leqslant \log\langle Y\rangle, \quad \text{(by Jensen)} \qquad (1) \\
&= \log\sum_{x\in\mathscr{A}} p(x)\frac{q(x)}{p(x)} \\
&= \log\sum_{x\in\mathscr{A}} q(x)
\end{aligned}$$

So,

$$\begin{aligned} -D(p\|q) &\leqslant \log \sum_{x\in\mathscr{A}} q(x) \\ &\leqslant \log \sum_{x\in\mathscr{X}} q(x) \qquad (2) \\ &= \log 1 = 0 \\ \implies D(p\|q) &\geqslant 0. \end{aligned}$$

log is strictly cap-like, so if (1) is an equality, then $\frac{p(x)}{q(x)} = c$ with probability 1. This means:

$$\begin{aligned} \sum_{x\in\mathscr{A}} q(x) = c \sum_{x\in\mathscr{A}} p(x) &= c \\ = \sum_{x\in\mathscr{X}} q(x) &= 1, \end{aligned}$$

if (2) is an equality. □

Corollary.

$$I(X;Y) \geqslant 0,$$

with equality $\iff X, Y$ *independent.*

Proof.

$$I(X;Y) = D\big(p_{X,Y}(x,y)\|p_X(x)p_Y(y)\big) \geqslant 0,$$

with $D\big(p_{X,Y}(x,y)\|p_X(x)p_Y(y)\big) = 0 \iff p_X, Y(x,y) = p_X(x)p_Y(y)$, i.e. X and Y are independent. □

Corollary.

$$D\big(p(y|x)\|q(y|x)\big) \geqslant 0,$$

with equality $\iff q(y|x) = p(y|x) \forall x, y.$

Proof. This follows directly from the information inequality with $p \to p(y|x)$ and $q \to q(y|x)$. □

Corollary.

$$I(X;Y|Z) \geqslant 0,$$

with $I(X;Y|Z) = 0 \iff X, Y$ *conditionally independent, i.e.* $p(x,y|z) = p(x|z)p(y|z) \forall x, y, z.$

Proof. This follows directly from the first corollary. □

3.4 Upper bound on information

Theorem.

$$H(X) \leqslant \log|\mathscr{X}|$$

where $|\mathscr{X}|$ *is the size of the set of outcomes, with equality* $\iff p(x) = 1/|\mathscr{X}|$*, i.e. a uniform distribution.*

Proof. We have X and $p(x)$. Define another distribution on $\mathscr{X}$ given by $u(x) = \frac{1}{|\mathscr{X}|}$. Then:

$$\begin{aligned} 0 \leqslant D(p\|u) &= \sum_{x\in\mathscr{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= -H(x) - \log\frac{1}{|\mathscr{X}|} \\ &= -H(X) + \log|\mathscr{X}|. \end{aligned}$$

The information inequality also gives that equality $\iff p(x) = u(x)\forall x$. □

Theorem.

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

with equality $\iff$ *X, Y independent.*

Proof.

$$0 \leqslant I(X;Y) = H(X) - H(X|Y).$$

□

Theorem.

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

with equality if $X_1, \ldots, X_n$ all independent.

Proof. For $n = 2$:

$$\begin{aligned} 0 \leqslant I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ \Rightarrow H(X) + H(Y) &\geqslant H(X,Y). \end{aligned}$$

Now use the chain rule:

$$\begin{aligned} H(X_1, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \\ &\leqslant \sum_{i=1}^{n} H(X_i). \end{aligned}$$

□

3.5 Log-sum inequality

Theorem. *For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,*

$$\sum a_i \log\frac{a_i}{b_i} \geqslant \left(\sum a_i\right)\log\frac{\sum a_i}{\sum b_i}$$

Proof. Assume without loss of generality that $a_i > 0,\ b_i > 0$. $f(t) = t\log t$ is cup-like (since $f''(t) > 0$). Define:

$$\alpha_i = \frac{b_i}{\sum_k b_k}, \quad t_i = \frac{a_i}{b_i},$$

and apply Jensen's inequality to $\mathcal{T} = \{t_i\}$, $p_T(t_i) = \alpha_i$, f:

$$\begin{aligned}
\langle f(T)\rangle &\geqslant f(\langle T\rangle) \\
\sum \alpha_i f(t_i) &\geqslant f\left(\sum \alpha_i t_i\right) \\
\sum \frac{b_i}{\sum_k b_k} f\left(\frac{a_i}{b_i}\right) &\geqslant f\left(\sum_i \frac{b_i}{\sum_k b_k}\frac{a_i}{b_i}\right) \\
\sum b_i f\left(\frac{a_i}{b_i}\right) &\geqslant \sum b_k f\left(\frac{\sum a_i}{\sum b_i}\right) \\
\sum a_i \log\frac{a_i}{b_i} &\geqslant \left(\sum a_i\right)\log\frac{\sum a_i}{\sum b_i}
\end{aligned}$$

□

3.6 Information inequality

Theorem.

$$D(p\|q) \geqslant 0.$$

Proof.

$$\begin{aligned}
D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&\geqslant \left(\sum_{x\in\mathcal{X}} p(x)\right)\log\frac{\sum p(x)}{\sum q(x)} \\
&= 1\log 1 = 0.
\end{aligned}$$

Equality $\iff \frac{p(x_i)}{q(x_i)} = \text{const.} \Rightarrow p(x_i) = q(x_i)$ for all i. □

3.7 Cup-likeness of KL divergence

If (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions, then:

$$(p_\lambda, q_\lambda) = \left(\lambda p_1 + (1-\lambda)p_2, \lambda q_1 + (1-\lambda)q_2\right)$$

is also a pair of probability distributions for $0 \leqslant \lambda \leqslant 1$.
Theorem. *For* (p_1, q_1) *and* (p_2, q_2), $\lambda \in [0,1]$,

$$D(p_\lambda, q_\lambda) \leqslant \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2).$$

Proof. Consider some $x \in \mathscr{X}$.

$$\big(\lambda p_1(x) + (1-\lambda)p_2(x)\big)\log\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \leqslant$$

$$\leqslant \lambda p_1(x)\log\frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}$$

by log-sum. Sum both sides over x:

$$D\big(\lambda p_1(x) + (1-\lambda)p_2(x)\|\lambda q_1(x) + (1-\lambda)q_2(x)\big)\lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)$$

□

3.8 Cap-likeness of entropy

Theorem.

$$H(\lambda p_1 + (1-\lambda)p_2) \geqslant \lambda H(p_1) + (1-\lambda)H(p_2).$$

Proof. X_1 and X_2 are two random variables on the same set of outcomes $\mathscr{X}$ with distributions p_1 and p_2; let θ be a random variable with two outcomes: $p_\theta(a) = \lambda$, $p_\theta(b) = 1-\lambda$. Consider the random process X_θ: $\theta = a \to X_1$, $\theta = b \to X_2$. So,

$$p_{X_\theta}(x) = \lambda p_1(x) + (1-\lambda)p_2(x).$$

$$\begin{aligned} H(X|\theta) &= \lambda H(X_1) + (1-\lambda)H(X_2) \\ &= \lambda H(p_1) + (1-\lambda)H(p_2) \\ &= -\sum p_1(x)\log p_1(x). \end{aligned}$$

We know $H(X) = H(\lambda p_1 + (1-\lambda)p_2)$, and $H(x) \geqslant H(X|\theta)$. So:

$$H(\lambda p_1 + (1-\lambda)p_2) \geqslant \lambda H(p_1) + (1-\lambda)H(p_2).$$

□

3.9 Another theorem

Theorem. *Let $(X,Y) \sim p(x,y) = p(x)p(y|x)$. Then the mutual information $I(X;Y)$ can be considered a function of p_X and $p_{Y|X}$, and is cup-like in $p(x)$ for fixed $p(y|x)$ and cap-like in $p(y|x)$ for fixed $p(x)$.*

Proof in book.

4 Markov chains & sufficient statistics

4.1 Markov chains

Random variables X, Y, Z are said to form a *Markov chain* denoted $X \to Y \to Z$ if $p(x,y,z) = p(x)p(y|x)p(z|y)$. Equivalently, X and Z are conditionally independent:

$$\begin{aligned} p(x,z|y) = \frac{p(x,y,z)}{p(y)} &= \frac{p(x,y)p(z|y)}{p(y)} \\ &= \frac{p(x|y)p(y)p(z|y)}{p(y)} \\ p(x,z|y) &= p(x|y)p(z|y) \end{aligned}$$

Note also that the definition is symmetric: $X \to Y \to Z \iff Z \to Y \to X$. Also, if $Z = f(Y)$ then $X \to Y \to Z$.

4.2 Data processing inequality

Theorem. *If $X \to Y \to Z$, then $I(X;Y) \geqslant I(X;Z)$.*

Proof.

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

Since $p(x,z|y) = p(x|y)p(z|y)$, then:

$$I(X;Z|Y) = \sum p(y) \sum \log \underbrace{\frac{p(x,z|y)}{p(x|y)p(z|y)}}_{=1} = 0$$

So $I(X;Y) = I(X;Z) + I(X;Y|Z) \geqslant I(X;Z)$, with equality if $I(X;Y|Z) = 0 \iff X \to Z \to Y$. □

Corollary. *If $Z = g(Y)$ then $I(X;Y) \geqslant I(X;g(Y))$.*

Proof. $X \to Y \to g(Y)$. □

Corollary. *If $X \to Y \to Z$, then $I(X;Y|Z) \leqslant I(X;Y)$.*

Proof. Recall $I(X;Y)+I(X;Z|Y) = I(X;Z)+I(X;Y|Z)$. Noting $I(X;Z|Y) = 0$ and $I(X;Z) \geqslant 0$ gives result. □

4.3 Sufficient statistics

$\{p_\theta(X)\}$ is a set of probability distributions indexed by some underlying variable θ. $T(X)$ is a statistic, e.g. mean, variance, mean and variance, ...

Since $T(X)$ is a function of X,

$$\theta \to X \to T(X),$$

so $I(\theta; X) \geqslant I(\theta; T(X))$.

$T(X)$ is a sufficient statistic if $I(\theta; X) = I(\theta; T(X))$. In other words, if:

$$\theta \to T(X) \to X.$$

For example, let $X_1, \ldots, X_n$ with $X_i \in \{0, 1\}$ be an independent and identically distributed (iid) set of random variables. Let $\theta = p_{X_i}(1)$. We want to estimate θ.

Let:

$$T(X) = \sum_{i=1}^{n} X_i.$$

Claim: $T(X)$ is a sufficient statistic. We show this by showing that X is independent of θ given $T(X)$.

$$\Pr\left((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \middle| \sum X_i = k\right) = \begin{cases} \frac{1}{N} & \text{if } \sum x_i = k \\ 0 & \text{otherwise,} \end{cases}$$

where $N = \binom{n}{k}$. This is clearly independent of θ. This means $\theta \to T(X) \to X$, so $T(X)$ is a sufficient statistic.

4.4 Fano inequality

X is a random variable. Y is a measurement. $\hat{X}$ is a guess of X. We are interested in the probability of error.

If $H(X|Y) \sim H(X)$, then reconstructing X from Y is just a guess – i.e. the chance of error is not different from guessing without knowing Y. If $H(X|Y) = 0$, then there is no uncertainty left in X when given Y, so Y determines X and there will be no error. The Fano inequality quantifies this.

X is a random variable; Y is another variable which is related to it. $\hat{X}$ is the reconstruction of X from Y; we usually imagine $\hat{X}$ to be a function of Y. Note that in general $\mathcal{X} \neq \hat{\mathcal{X}}$. Denote the probability of error $p_e = \Pr(X \neq \hat{X})$.

Theorem. *If $\hat{X}$ is an estimator: $X \to Y \to \hat{X}$, and $p_e = \Pr(X \neq \hat{X})$, then:*

$$\begin{aligned} H(p_e) + p_e \log|\mathcal{X}| &\geqslant H(X|\hat{X}) \\ &\geqslant H(X|Y). \end{aligned}$$

This inequality is often weakened to:

$$1 + p_e \log|\mathscr{X}| \geqslant H(X|Y),$$

(since $H(p_e) \geqslant 1$) or:

$$p_e \geqslant \frac{H(X|Y) - 1}{\log|\mathscr{X}|}.$$

Proof.

$$H(X|\hat{X}) \geqslant H(X|Y)$$

since $I(X;\hat{X}) \leqslant I(X;Y)$. So:

$$H(X) - H(X|\hat{X}) \leqslant H(X) - H(X|Y).$$

Define:

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$$

so that $p_E(1) = p_e$.

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(X|\hat{X}), \end{aligned}$$

since E is known given X and $\hat{X}$.

$$H(E, X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}).$$

We know:

$$H(E|\hat{X}) \leqslant H(E) = H(p_e)$$

and

$$\begin{aligned} H(X|E, \hat{X}) &= p_E(0) H(X|\hat{X}, E = 0) + p_E(1) H(X|\hat{X}, E = 1) \\ &\leqslant 0 + \log|\mathscr{X}| p_e, \end{aligned}$$

since $X = \hat{X}$ when $E = 0$, Putting this together gives the Fano inequality:

$$H(p_e) + p_e \log|\mathscr{X}| \geqslant H(X|Y)$$

□

Corollary. *If $\hat{\mathscr{X}} \subseteq \mathscr{X}$ then the inequality can be tightened to:*

$$H(p_e) + p_e \log(|\mathscr{X}| - 1) \geqslant H(X|\hat{X}).$$

Proof. Before, we had $H(X|\hat{X}, E = 1) \leqslant \log|\mathscr{X}|$. However if $\hat{X} \in \mathscr{X}$ and we know $E = 1$, then X can only take $|\mathscr{X}| - 1$ different values. □

Corollary. *For any two random variables X and Y, let $p = \Pr(X \neq Y)$. Then $H(p) + p \log|\mathscr{X}| \geqslant H(X|Y)$.*

Proof. Simply apply the Fano inequality to the Markov chain $X \to Y \to Y$. □

5 Asymptotic equipartition principle & typical sets

5.1 Convergence

Convergence: given a sequence of random variables $Y_1, Y_2, \ldots$ we say the sequence converges to Y:

- "in probability" if $\forall \varepsilon > 0$, $\Pr(|Y_n - Y| > \varepsilon) \to 0$
- "in mean square" if $\left\langle (Y - Y_n)^2 \right\rangle \to 0$.
- "with probability 1" or "almost surely" if:

$$\Pr\left(\lim_{n\to\infty} Y_n = Y\right) = 1$$

5.2 Law of large numbers

If $X_1, X_2, \ldots$ is a sequence of iid random variables with $X_i \sim X$, $\mu = \langle X \rangle$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,

- **Weak law:** $\overline{X}_n \to \mu$ in probability.
- **Strong law:** $\overline{X}_n \to \mu$ almost surely.

The asymptotic equipartition principle is basically the law of large numbers for information theory.

5.3 Asymptotic equipartition principle

The AEP states that:

$$p(X_1, X_2, \ldots, X_n) \sim 2^{-nH(X)}$$

for iid random variables $X_1, \ldots, X_n$ with $X_i \sim X$.

For example, if $\mathscr{X} = \{0, 1\}$ with probabilities p and q respectively, one possible sequence is $(1, 1, \ldots, 1)$, but it is not the most likely: it has probability q^n.

Usually the sequence will consist of $\sim np$ 0s and $\sim nq$ 1s. Such a sequence has probability $p^{np}q^{nq}$, and:

$$\begin{aligned} p^{np}q^{nq} &= 2^{\log p^{np}} 2^{\log q^{nq}} \\ &= 2^{\log p^{np} + \log q^{nq}} \\ &= 2^{-nH}. \end{aligned}$$

So 2^{-nH} is, roughly speaking, the probability of a typical outcome.

Theorem. *If $X_1, X_2, \ldots, X_n$ are iid with $X_i \sim X$, then:*

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

Proof. Functions of independent variables are independent: Y independent of $Z \Rightarrow f(Y)$ independent of $f(Z)$. So $\{\log p(X_i)\}$ are independent variables. By the weak law of large numbers,

$$\frac{1}{N}\sum \log p(X_i) \to \langle \log p(X) \rangle$$

The theorem follows from noting that $\langle \log p(x) \rangle = -H(X)$, and:

$$\begin{aligned}\sum \log p(X_i) &= \log \prod p(X_i) \\ &= \log p(X_1, \ldots, X_n).\end{aligned}$$

□

5.4 Typical set

The typical set $A_\varepsilon^{(n)}$ is the set of elements $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ such that:

$$2^{-n(H(X)+\varepsilon)} \leqslant p(x_1, \ldots, x_n) \leqslant 2^{-n(H(X)-\varepsilon)}$$

We saw that $p(X_1, \ldots, X_n) \to 2^{-nH(X)}$, so $A_\varepsilon^{(n)}$ consists of the $(x_1, \ldots, x_n)$ which are close to the limit.

Theorem. *1. If $(x_1, \ldots, x_n) \in A_\varepsilon^{(n)}$, then:*

$$H(X) - \varepsilon \leqslant -\frac{1}{n}\log p(x_1, \ldots, x_n) \leqslant H(X) + \varepsilon$$

2. $\Pr(A_\varepsilon^{(n)}) > 1 - \varepsilon$ for sufficiently large n.

3. $|A_\varepsilon^{(n)}| \leqslant 2^{n(H(X)+\varepsilon)}$, and:

4. $|A_\varepsilon^{(n)}| \geqslant 2^{n(H(X)-\varepsilon)}$.

Proof. 1. Follows from taking the log of the definition of the typical set.

2. We know:

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

In other words:

$$\Pr\left(\left|-\tfrac{1}{n}\log p(X_1, \ldots, X_n)\right| < \varepsilon\right) \to 1.$$

or in terms of δ,

$$\Pr\left(\left|-\tfrac{1}{n}\log p(X_1, \ldots, X_n)\right| < \varepsilon\right) > 1 - \delta$$

and let $\delta = \epsilon$ for result.

3.

$$1 = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \geqslant \sum_{\mathbf{x} \in A_\varepsilon^{(n)}} p(\mathbf{x}) \geqslant \sum_{\mathbf{x} \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)}.$$

So:

$$\begin{aligned}1 &\geqslant 2^{-n(H(X)+\varepsilon)} \sum_{\mathbf{x} \in A_\varepsilon^{(n)}} 1 \\ &= 2^{-n(H(X)+\varepsilon)} \left|A_\varepsilon^{(n)}\right|.\end{aligned}$$

4.

$$\begin{aligned}1-\varepsilon &< \Pr(A_\varepsilon^{(n)}) \\ &= \sum_{\mathbf{x}\in A_\varepsilon^{(n)}} p(\mathbf{x}) \\ &\leqslant \sum_{\mathbf{x}\in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} \\ &= \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X)-\varepsilon)}.\end{aligned}$$

□

5.5 Code word length

Let $X_1, X_2, \ldots$ be iid with probability distribution $p(x)$, and we wish to find a short description of a length-n output. We can construct a code word by letting the first bit be `0` if $\mathbf{x} \in A_\varepsilon^{(n)}$ or `1` otherwise. Then assign some order to elements of both $A_\varepsilon^{(n)}$ and $\mathscr{X}^n \setminus A_\varepsilon^{(n)}$, convert the number of the element to binary and concatenate.

For example, the 18^{th} element of the typical set would have a code word `0 10010` by this scheme.

The length of this binary number must be less than $\log|\text{set}| + 1$. Hence the elements of the typical set have code words of length $l < \log 2^{n(H+\varepsilon)} + 1 + 1 = n(H+\varepsilon) + 2$. Similarly words not in the typical set have length $l < \log|\mathscr{X}^n \setminus A_\varepsilon^{(n)}| + 2$, $\Rightarrow l < \log|\mathscr{X}^n| + 2$.

What is the average length of a code word?

$$\begin{aligned}\langle l(\mathbf{x})\rangle &= \sum p(\mathbf{x})\, l(\mathbf{x}) \\ &= \sum_{x\in A_\varepsilon^{(n)}} p(\mathbf{x})\, l(\mathbf{x}) + \sum_{x\notin A_\varepsilon^{(n)}} p(\mathbf{x})\, l(\mathbf{x}) \\ &\leqslant \sum_{x\in A_\varepsilon^{(n)}} p(\mathbf{x})\big[n(H+\varepsilon)+2\big] + \sum_{x\notin A_\varepsilon^{(n)}} p(\mathbf{x})\big[\log|\mathscr{X}^n|+2\big] \\ &= \Pr(A_\varepsilon^{(n)})\big[n(H+\varepsilon)+2\big] + \Pr(\mathscr{X}^n \setminus A_\varepsilon^{(n)}) \\ &\leqslant (1-\varepsilon)\big[n(H+\varepsilon)+2\big] + \varepsilon\big[\log|\mathscr{X}^n|+2\big] \\ &= n(H+\varepsilon'),\end{aligned}$$

where $\varepsilon' \equiv \varepsilon + \varepsilon\log|\mathscr{X}| + \frac{2}{n}$. So:

$$\frac{\langle l(\mathbf{x})\rangle}{n} \leqslant H + \varepsilon'.$$

ε' can be made arbitrarily small for sufficiently large n and suitable choice of ϵ. This means that for iid $X_1, \ldots, X_n$ with $X_i \sim X$, the code length per outcome can be made arbitrarily close to H.

6 Encoding

6.1 Definitions

Ideally we want common outcomes to have short code words, and rare outcomes to have long code words. For example, in Morse code, e is ·, t is –, a is · –, etc.

We define a *source code* C for a variable X as a mapping:

$$C : \mathscr{X} \to \mathscr{D}^*, \quad x \mapsto C(x)$$

where $\mathscr{D}^*$ is the set of finite-length strings of elements from the D-symbol alphabet $\mathscr{D}$. The length of $C(x)$ is denoted $l(x)$.

The *expected length* of a source code, $L(C)$ is the average length:

$$\begin{aligned} L(C) &= \langle l(x) \rangle \\ &= \sum_{x \in \mathscr{X}} p(x)\, l(x). \end{aligned}$$

We say a code is *non-singular* if every $x \in \mathscr{X}$ maps to a different string:

$$x \neq x' \Rightarrow C(x) \neq C(x').$$

i.e. c is injective.

The *extension* C^* of a code C is the induced map:

$$C^* : \mathscr{X}^* \to \mathscr{D}^*$$

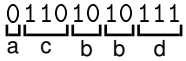
given by concatenation:

$$C(x_1 x_2 \ldots x_n) = C(x_1)\, C(x_2) \ldots C(x_n).$$

A code is *instantaneous* (or a *prefix code*) if no code word is the prefix of another, e.g.:

a	b	c	d
0	10	110	111

A prefix code (e.g. 01101010111) can be uniquely split up:



However, codes can have this property without being prefix codes.

6.2 Kraft inequality

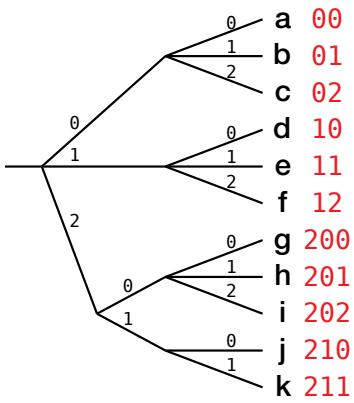
Theorem. *For any prefix code over an alphabet of size D with code word lengths $l_1 = l(x_1), l_2, \ldots, l_n$,*

$$\sum_{i=1}^{n} D^{-l_i} \leqslant 1.$$

Conversely, given $\{l_i\}$ satisfying this inequality it is possible to construct a prefix code.

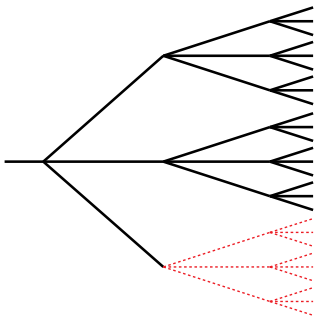
Note that this inequality has nothing to do with efficiency – it is about prefix-ness.

Proof. Construct a tree where code words give a 'route map'. Each node in the tree has $\leqslant D$ outgoing connections labelled $0, \ldots, D-1$. For example, here outcomes a – k are assigned code words from a 3-nary alphabet:



This example is a prefix code, since all of the outcomes correspond to *leaves*: nodes with no outgoing connections. Leaves have no *descendants*: nodes which a given node prefixes.

Call the length of the longest code word l_* for a given prefix code. If all leaves at this level were in the code, we would have D^{l_*} leaves. If there is a code word of length $l_i < l_*$, then it would have had $D_{l_*-l_i}$ descendants at length l_*. For example, here a code word of length 1 would have had $3^{3-1} = 9$ descendants of length 3.



Because this is a prefix code, no two code words share any descendants. So the number of children at level l_* must be less than or equal to the maximum number of children at

level l_*:

$$\sum_{i=1}^{n} D^{l_*-l_i} \leqslant D^{l_*}.$$

Dividing by D^{l_*} gives the Kraft inequality.

Conversely if $l_1,\ldots,l_n$ satisfy the Kraft inequality, we can choose disjoint family subsets of size $D^{l_*-l_i}$ at level l_* and cut them off at their parent to form a prefix code. □

Note that this is for $|\mathscr{X}|$ finite. There does exist an extension for countably infinite $\mathscr{X}$.

6.3 Optimal codes

The idea is to find a lower bound on $L(C)$. We can do this by minimising $L(C) = \sum p_i l_i$ subject to $\sum D^{-l_i} \leqslant 1$. Use Lagrange multipliers!

$$J = \sum p_i l_i + \lambda\left(\sum D^{-l_i} - 1\right).$$

Now minimise J:

$$\begin{aligned}\frac{\partial J}{\partial l_i} &= p_i - \lambda D^{-l} \ln D_i = 0\\ \Rightarrow D^{-l_i} &= \frac{p_i}{\lambda \ln D}\end{aligned}$$

Sum both sides over i:

$$1 = \frac{1}{\lambda \ln D}$$

and substitute back in:

$$D^{-l_i} = p_i.$$

So the optimal length, l_i^*, is given by $-\log_D p_i$. The average length of a code word is given by:

$$\begin{aligned}L^* &= \sum p_i l_i^*\\ &= -\sum p_i \log_D p_i\\ &= H_D(X).\end{aligned}$$

Note that $-\log_D p_i$ is in general not an integer. We say that the best code has $l_i = \lceil l_i^* \rceil$, i.e. l_i^* rounded up to the nearest integer.

The *Shannon–Fano theorem* states that for an optimal prefix code:

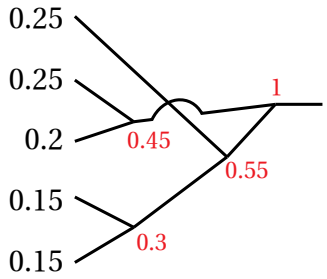
$$H_D \leqslant L \leqslant H_D + 1.$$

In other words, the average code word length of an optimal prefix code is within 1 digit of the minimum possible average length.

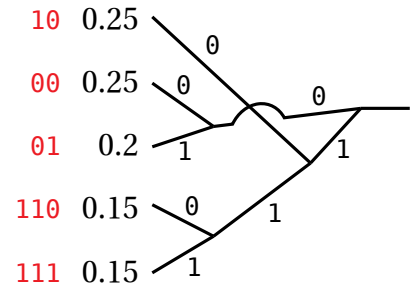
6.4 Huffman coding

An elegant scheme for constructing an optimal prefix code.

Begin with the outcomes as leaves labelled by their probabilities. Then repeatedly join the D nodes with the lowest probabilities (sometimes this can be done in more than one way) and label the new node with the sum of these probabilities. Here this is performed for $D = 2$:

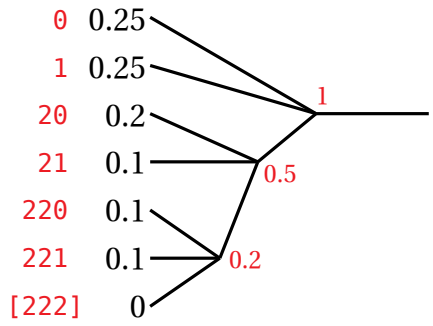


Labelling routes yields code words for each outcome:



This gives a prefix code where the most probable outcomes have the shortest code words.

For $D \geqslant 3$ there may not be the correct number of symbols. Since each joining replaces D symbols with 1, and we want 1 left at the end, this alogrithm only works for numbers of the form $k(D-1)+1$. To solve this, introduce a sufficient number of 0-probability nodes. For example, for $D = 3$:



This means that there will be codes that do not correspond to any outcomes (for example, 222 above). This means the code is in some sense not as efficient as it 'could be'. However the code will still have $L \leqslant H_D + 1$.