Why entropy represents information?

* A source code C for a random variable X is a mapping from the range of X to the set of finite-length strings of symbols from a D-ary alphabet, denote $\mathcal{D}^*$.

  $C(x)$ — code word correspond to x

  $l(x)$ — length of $C(x)$

- Binary alphabet. $\{0, 1\}$
- A code is called a <u>prefix code</u> or an <u>instantaneous code</u> if no codeword is a prefix of any other code-word.

$p(x)$	X	not instantaneous	instantaneous
$\frac{1}{2}$	1	10	0
$\frac{1}{4}$	2	00	10
$\frac{1}{8}$	3	11	110
$\frac{1}{8}$	4	110	111

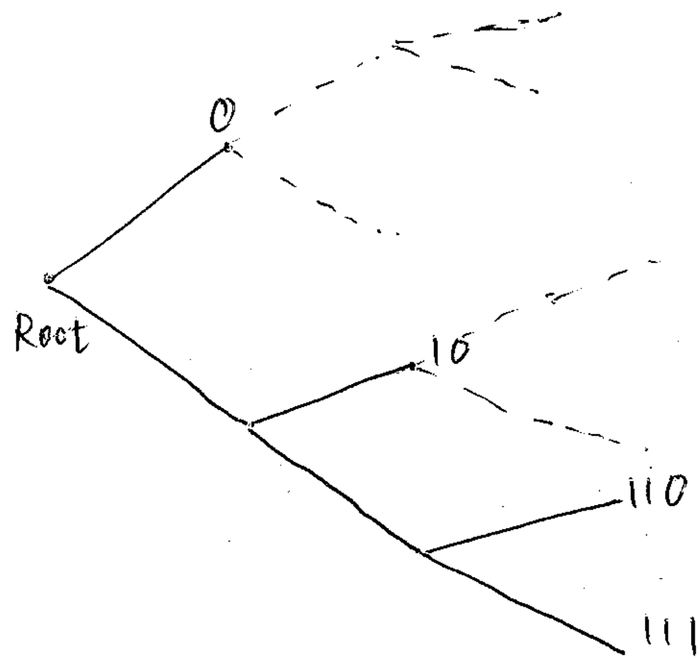
- Expected length of a source code $C(x)$ for a r.v. X with prob. mass $p(x)$ is

$$L(C) = \sum_x p(x)\, l(x)$$

- Kraft Inequality:

  Instantaneous code over alphabet of size D $\iff \sum_i D^{-l_i} \le 1$

Let's look at an example for binary alphabet. The code tree ~~must~~ for an instantaneous code looks like



l_{max} — the length of the longest code word.

A codeword at level ~~l_{max}~~ l_i has $D^{l_{max}-l_i}$ decendants at level l_{max}.

$$\sum_{i} D^{l_{max}-l_i} \leq D^{l_{max}}$$

$$\Rightarrow \sum_i D^{-l_i} \leq 1$$

Optimal Codes

$$\min \ L = \sum_i p_i l_i$$

$$\text{s.t.} \ \sum_i D^{-l_i} \leq 1$$

l_i integers

Let me go sloppy here.

Drop the integrality constraint, and assume that

$\sum_i D^{-l_i} = 1$ at optimal.

Then, ~~the~~ we have

$$G = \sum_i P_i l_i + \lambda \left(\sum D^{-l_i} - 1 \right)$$

$$\frac{\partial G}{\partial l_i} = P_i - \lambda D^{-l_i} \cdot \log_e D = 0 \qquad \left(D^{-l_i} = e^{-l_i \cdot \log_e D} \right)$$

$$D^{-l_i} = \frac{P_i}{\lambda \log_e D}$$

$$\sum D^{-l_i} = 1 \quad \Rightarrow \lambda = 1/\log_e D$$

$$\Rightarrow P_i = D^{-l_i}$$

$$\Rightarrow l_i^* = -\log_D P_i.$$

It can be verified that this is a global minimum for the LP relaxation. (Omit the proof here.)

$$\therefore \quad L \geq \sum_i P_i l_i^* = -\sum_i P_i \log_D P_i = H_D(X)$$

for the optimal code.

Note that we can chose

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$$

This satisfies the Kraft inequality.

$$\sum D^{-\lceil \log \frac{1}{p_i} \rceil} \leq \sum D^{-\log \frac{1}{p_i}} = \sum p_i = 1$$

$$\log_D \frac{1}{p_i} \leq l_i < \log_D \frac{1}{p_i} + 1$$

$$\Rightarrow \quad H_D(X) \leq L < H_D(X) + 1$$

This is the bound for optimal code length.

* differential entropy for continuous probability distribution.

$h(X) = -\int_S f(x) \log f(x) \cdot dx$ where S is the support of X

Not all properties of entropy hold.

E.g. $f(x)$ can be greater than 1.

$$X \sim U(0, \tfrac{1}{2})$$

$$h(x) = -\int_0^{\frac{1}{2}} 2 \log 2 \, dx = -\log 2 < 0$$