Lecture Notes for Information Theory (ECE 587/STA 563)

Jyunyi Liao

Contents

1 Measure of Information

Throughout this section, we assume that all random variables we study are discrete variables. We use capital letters like X, Y, Z to denote random variables, and their probability mass functions $p_X(x), p_Y(y), p_Z(z)$. For simplicity, we drop the subscripts and use the shorthand $p(x)$, $p(y)$, $p(z)$ instead. We use calligraphy letters like $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to denote the finite support of random variables.

1.1 Entropy and Conditional Entropy

Definition 1.1 (Entropy)**.** *Let X be a random variable supported on a finite state space $\mathcal{X}$, with probability mass function $p(x)$. The entropy of X is a function of the distribution $p(x)$:*

$$H(X) := \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = -\mathbb{E}\left[\log p(X)\right].$$

Likewise, for a collection $X_1, \cdots, X_n$ of random variables, the (joint) entropy of $X_1, \cdots, X_n$ is defined as the entropy of the random vector $(X_1, \cdots, X_n)$:

$$H(X_1, \cdots, X_n) = \sum_{x_1\in\mathcal{X}_1, \cdots, x_n\in\mathcal{X}_n} p(x_1, \cdots, x_n) \log \frac{1}{p(x_1, \cdots, x_n)}.$$

Remark I. The entropy provides a measure of uncertainty of random variables. We also frequently use the *binary entropy function* $h : [0, 1] \to \mathbb{R}_+$, which is defined as the entropy of a Bernoulli variable:

$$h(\alpha) = H(\text{Bernoulli}(\alpha)) = -\alpha \log \alpha - (1-\alpha) \log(1-\alpha), \quad \alpha \in [0, 1]$$

with the convention $0 \log 0 = 0$.

Remark II. Given any base $b > 0$, we define the entropy of X under base b to be

$$H_b(X) = \sum_{x\in\mathcal{X}} p(x) \log_b \frac{1}{p(x)} = H(X) \log_b \mathrm{e}.$$

Clearly we have $H(X) = H_{\mathrm{e}}(X)$. Another commonly used entropy is the bit entropy, in which the base $b = 2$:

$$H_2(X) = \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = H(X) \log_2 \mathrm{e}.$$

Proposition 1.2. *We have the following estimate for the entropy of a random variable X:*

$$0 \le H(X) \le \log |\mathcal{X}|.$$

Proof. The lower bound follows from the definition of entropy. For the upper bound, note that

$$\begin{aligned}\sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = \sum_{x\in\mathcal{X}} p(x) \log \frac{|\mathcal{X}|}{p(x)|\mathcal{X}|} &= \log |\mathcal{X}| + \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)|\mathcal{X}|} \\ &\le \log |\mathcal{X}| + \sum_{x\in\mathcal{X}} p(x) \left(\frac{1}{p(x)|\mathcal{X}|} - 1\right) = \log |\mathcal{X}|.\end{aligned}$$

Then we complete the proof. □

Remark. If $|\mathcal{X}| = \infty$, the entropy of a random variable can be ∞. For example, let $A = \sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$, which is less than infinity. Define random variable X by

$$\mathbb{P}(X = n) = \frac{1}{An(\log n)^2}, \quad n = 2, 3, \cdots.$$

Then

$$H(X) \geq \int_2^{\infty} \frac{\log A}{x \log x} \, dx = \infty.$$

We may also wonder the uncertainty of a random variable when given potentially relevant observation.

Definition 1.3 (Conditional Entropy)**.** *Let X and Y be two random variables in the same probability space. The entropy of Y conditioned on the event $X = x$ is a function of the conditional distribution $p(y|x)$:*

$$H(Y|X = x) := \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} = \mathbb{E}\left[\log \frac{1}{p(Y|x)} \,\middle|\, X = x \right].$$

The conditional entropy of Y given X is a function of the joint distribution $p(x, y)$:

$$H(Y|X) := \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} = \mathbb{E}\left[\log \frac{1}{p(Y|X)} \right].$$

Remark. Note that $H(Y|X)$ is a deterministic quantity rather than a random variable. In fact, we have

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

Next, we study the relation between joint entropy and conditional entropy.

Proposition 1.4 (Chain rule for entropy)**.** *The joint entropy of X and Y has the following decomposition:*

$$H(X, Y) = H(Y|X) + H(X). \tag{1.1}$$

More generally,

$$H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \cdots, X_1). \tag{1.2}$$

Proof. We first verify the bivariate case (1.1):

$$\begin{aligned}
H(Y|X) + H(X) &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} + \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x)} \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} = H(X, Y).
\end{aligned}$$

The general case (1.2) follows from mathematical induction. □

Remark. The equality (1.1) also implies the chain rule for conditional entropy:

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

Finally, we introduce an important property of entropy as the function of distribution.

Theorem 1.5 (Concavity of entropy)**.** *Let p and q be two probability distributions that are supported in a common space $\mathcal{X}$. Then for all $0 \le \lambda \le 1$, we have*

$$H(\lambda p + (1-\lambda)q) \ge \lambda H(p) + (1-\lambda)H(q). \tag{1.3}$$

Proof. We simply employ the estimate $\log t \le t - 1$ on $\lambda H(p) + (1-\lambda)H(q) - H(\lambda p + (1-\lambda)q)$:

$$\begin{aligned}
&\lambda \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} + (1-\lambda) \sum_{x\in\mathcal{X}} q(x) \log \frac{1}{q(x)} - \sum_{x\in\mathcal{X}} (\lambda p(x) + (1-\lambda)q(x)) \log \frac{1}{\lambda p(x) + (1-\lambda)q(x)} \\
&= \lambda \sum_{x\in\mathcal{X}} p(x) \log \frac{\lambda p(x) + (1-\lambda)q(x)}{p(x)} + (1-\lambda) \sum_{x\in\mathcal{X}} q(x) \log \frac{\lambda p(x) + (1-\lambda)q(x)}{q(x)} \\
&\le \lambda \sum_{x\in\mathcal{X}} (\lambda p(x) + (1-\lambda)q(x) - p(x)) + (1-\lambda) \sum_{x\in\mathcal{X}} (\lambda p(x) + (1-\lambda)q(x) - q(x)) = 0.
\end{aligned}$$

Then the result follows. $\square$

Remark. Using the concavity, we can interpret why a transfer of probability that makes the distribution more uniform increases the entropy. We consider the following transformation:

$$(p_1, \cdots, p_i, \cdots, p_j, \cdots, p_m) \to \left(p_1, \cdots, \frac{p_i + p_j}{2}, \cdots, \frac{p_i + p_j}{2}, \cdots, p_m\right), \quad p_1 + \cdots + p_m = 1.$$

Let $p = (p_1, \cdots, p_i, \cdots, p_j, \cdots, p_m)$, and let $q = (p_1, \cdots, p_j, \cdots, p_i, \cdots, p_m)$ be the probability vector with i-th and j-th elements exchanged. Then

$$H\left(\frac{p+q}{2}\right) \ge \frac{1}{2}H(p) + \frac{1}{2}H(q) = H(p).$$

1.2 Mutual Information

Definition 1.6 (Mutual information)**.** *Let X and Y be two discrete random variables in the same probability space. The mutual information of X and Y is defined as*

$$I(X;Y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

Proposition 1.7 (Properties of mutual information)**.** *Let X and Y be two discrete random variables.*

(i) (Symmetry). $I(X;Y) = I(Y;X)$.

(ii) (Reduction). $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

(iii) (Measure of dependency). $I(X;Y) \ge 0$, and the equality holds if and only if X and Y are independent.

Proof. The assertion (i) follows from definition, and the second from direct calculation. Now we verify (iii):

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ge \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \left(1 - \frac{p(x)p(y)}{p(x,y)}\right) = 0.$$

Clearly, the equality holds if and only if $p(x,y) = p(x)p(y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. $\square$

Remark. Combining (ii) and (iii), we know that *conditioning does not increase entropy*:

$$H(X|Y) \le H(X), \quad \textit{and} \quad H(Y|X) \le H(Y).$$

An alternative proof of Theorem 1.5. Let $X_1 \sim p$ and $X_2 \sim q$ be two independent random variables, and let $Z \sim \text{Bernoulli}(\lambda)$. Define

$$Y = X_1 \mathbb{1}_{\{Z=1\}} + X_2 \mathbb{1}_{\{Z=0\}}.$$

Then $Y \sim \lambda p + (1-\lambda)q$, and

$$H(Y) \geq H(Y|Z) = \lambda H(Y|Z=1) + (1-\lambda) H(Y|Z=0) = \lambda H(X_1) + (1-\lambda) H(X_2).$$

This is in fact the equality (1.3).

Definition 1.8. *Let X, Y and Z be discrete random variables in the same probability space. The conditional mutual information of X and Y given Z is defined as*

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

Similar to Proposition 1.7, conditional mutual information has the following properties.

Proposition 1.9 (Properties of conditional mutual information)**.** *Let X, Y and Z be discrete random variables in the same probability space.*

(i) (Symmetry). $I(X;Y|Z) = I(Y;X|Z)$.

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

(iii) (Measure of dependency). $I(X;Y|Z) \geq 0$, and the equality holds if and only if X and Y are conditionally independent on Z.

By direct calculation and induction, we also have the following chain rule for mutual information.

Proposition 1.10 (Chain rule for mutual information)**.** *The mutual information $I(X;Y,Z)$ has the following decomposition:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z).$$

More generally,

$$I(X;Y_1,Y_2,\cdots,Y_n) = I(X;Y_1) + I(X;Y_2|Y_1) + I(X;Y_3|Y_2,Y_1) \cdots + I(X;Y_n|Y_{n-1},\cdots,Y_1).$$

We can use this rule to derive the data processing inequality for Markov chains.

Definition 1.11 (Markov chain)**.** *Random variables X, Y and Z are said to form a Markov chain, written $X \to Y \to Z$, if X and Z are conditionally independent on Y:*

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

Particularly, if $Z = g(Y)$ is a function of Y, then $X \to Y \to Z$.

The following theorem asserts that no manipulation of Y can increase the mutual information.

Theorem 1.12 (Data processing inequality)**.** *If $X \to Y \to Z$, then*

$$I(X;Y) \geq I(X;Z).$$

Particularly, for any function g defined on $\mathcal{Y}$, we have

$$I(X;Y) \geq I(X;g(Y)).$$

Proof. By chain rule, we have that

$$I(X;Y) + I(X;Z|Y) = I(X;Y,Z) = I(X;Z) + I(X;Y|Z).$$

Since $X \perp\!\!\!\perp Z \,|\, Y$, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, the result follows. $\square$

Remark. By Proposition 1.7, we also have $H(X|Z) \geq H(X|Y)$ when $X \to Y \to Z$.

Next, we introduce an alternative definition of mutual information.

Definition 1.13 (Kullback-Leibler divergence/relative entropy)**.** *Let p and q be two probability distributions such that $\mathcal{X} = \operatorname{supp} q \supset \operatorname{supp} p$. The Kullback-Leibler divergence of q from p is defined as*

$$D(p\|q) := \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right].$$

This is also known as the relative entropy.

Remark. By definition, we have

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \geq \sum_{x\in\mathcal{X}} p(x) \left(1 - \frac{q(x)}{p(x)}\right) = 0.$$

Therefore, $D(p\|q) \geq 0$, and the equality holds if and only if $p = q$. Moreover, by definition, we have the following result:

$$I(X;Y) = D(p_{X,Y}\|p_X p_Y) = \mathbb{E}_{X\sim p_X}\left[D(p_{Y|X}\|p_Y)\right].$$

In other words, the mutual information of X and Y is the relative entropy of their marginal product $p_X p_Y$ from their joint distribution $p_{X,Y}$.

Application: Misclassification Rate. To end this section, we introduce a useful application of mutual information. We discuss the estimation of a discrete random variable X from an observation Y. To deal with this problem, we construct a function $\phi : \mathcal{Y} \to \mathcal{X}$. The probability of error of the estimator $\widehat{X} = \phi(Y)$ is

$$p_e = \mathbb{P}(\widehat{X} \neq X).$$

The following Fano's inequality provide a lower bound of the error rate p_e.

Theorem 1.14 (Fano's inequality)**.** *For any estimator $\widehat{X}$ of X such that $X \to Y \to \widehat{X}$, we have*

$$p_e \geq \frac{H(X|Y) - \log 2}{\log|\mathcal{X}|}.$$

Proof. Let $B = \mathbb{1}_{\{X=\widehat{X}\}}$, which is a Bernoulli variable with parameter p_e. By the chain rule, the conditional entropy of (B, X) given $\widehat{X}$ is

$$H(B|\widehat{X}) + H(X|B,\widehat{X}) = H(B,X|\widehat{X}) = H(X|\widehat{X}) + H(B|X,\widehat{X}).$$

Now we analyze the four terms in the equality.

(i) Since conditioning does not increase entropy, $H(B|\widehat{X}) \le H(B) = h(p_e)$.

(ii) The conditional entropy $H(X|B,\widehat{X})$ has the following estimate:

$$\begin{aligned}
H(X|B,\widehat{X}) &= \sum_{b\in\{0,1\}} \sum_{x\in\mathcal{X}} \sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=b, X=x, \widehat{X}=\widehat{x}) \log \frac{1}{\mathbb{P}(X=x|B=b,\widehat{X}=\widehat{x})} \\
&= \sum_{x\in\mathcal{X}} \sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=0, X=x, \widehat{X}=\widehat{x}) \log \frac{1}{\mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x})} \\
&= \sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=0, \widehat{X}=\widehat{x}) \underbrace{\sum_{x\in\mathcal{X}} \mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x}) \log \frac{1}{\mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x})}}_{\le \log|\mathcal{X}|} \\
&\le p_e \log|\mathcal{X}|.
\end{aligned}$$

(iii) Since $X \to Y \to \widehat{X}$, the data processing inequality implies $H(X|\widehat{X}) \ge H(X|Y)$.

(iv) Since B is a function of X and $\widehat{X}$, we have $H(B|X,\widehat{X}) = 0$.

Combining these estimates, we obtain

$$H(X|Y) \le h(p_e) + p_e \log|\mathcal{X}| \le \log 2 + p_e \log|\mathcal{X}|.$$

Then we complete the proof. $\square$

1.3 The Typical Set and Asymptotic Equipartition Property

In this section, we investigate a sequence of i.i.d. copies $X_1, X_2, \cdots$ of a random variable $X \sim p(x)$ with finite support $\mathcal{X}$. We write for a random vector of length n and its realization

$$X_{1:n} = (X_1, \cdots, X_n), \quad x_{1:n} = (x_1, \cdots, x_n).$$

The joint distribution of $X_{1:n}$ is given by

$$p(x_{1:n}) = \mathbb{P}\left(X_{1:n} = x_{1:n}\right) = \prod_{i=1}^{n} p(x_i).$$

In this section, our key task is to find a confidence set $A \subset \mathcal{X}^n$ that contains our observation $X_{1:n}$ with a high probability. Formally, we require

$$\mathbb{P}(X_{1:n} \in A) \ge 1 - \epsilon,$$

where $\epsilon > 0$ is an arbitrarily given small quantity.

Typical Sets. We first propose an idea of constructing high probability sets. Let $g : \mathcal{X} \to \mathbb{R}$ be a function such that $\mathbb{E}|g(X)| < \infty$. By the weak law of large number, for each $\epsilon > 0$, there exists $N_\epsilon > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)]\right| \le \epsilon\right) \ge 1 - \epsilon \quad \forall n \ge N_\epsilon.$$

Consequently, almost all probability mass is concentrated on the following set A:

$$A = \left\{x_{1:n} \in \mathcal{X}^n : \mathbb{E}\left[g(X)\right] - \epsilon \le \frac{1}{n}\sum_{i=1}^{n} g(x_i) \le \mathbb{E}\left[g(X)\right] + \epsilon\right\}.$$

In the last display, the constraint can be equivalently expressed as

$$2^{-n(\mathbb{E}[g(X)]+\epsilon)} \le 2^{-\sum_{i=1}^{n} g(x_i)} \le 2^{-n(\mathbb{E}[g(X)]-\epsilon)}.$$

The construction of typical sets follows by plugging in $g(x) = \log_2 \frac{1}{p(x)}$.

Definition 1.15. *The ϵ-typical set is defined by*

$$A_\epsilon^{(n)} = \left\{ x_{1:n} \in \mathcal{X}^n : 2^{-n(H_2(X)+\epsilon)} \le p(x_{1:n}) \le 2^{-n(H_2(X)-\epsilon)} \right\},$$

or equivalently, the set of all tuples $x_{1:n} \in \mathcal{X}^n$ obeying

$$H_2(X) - \epsilon \le -\frac{1}{n} \log_2 p(x_{1:n}) \le H_2(X) + \epsilon.$$

Clearly, there exists N_ϵ such that $A_\epsilon^{(n)}$ contains $X_{1:n}$ with probability at least $1-\epsilon$ whenever $n > N_\epsilon$.

Size of Typical Sets. When n increased, the number of possible realizations of $X_{1:n}$ would rise very quickly, which is $|\mathcal{X}|^n$. The idea of typical sets is to concentrate the probability mass of $X_{1:n}$ on a smaller set $A_\epsilon^{(n)}$:

$$A_\epsilon^{(n)} = \left\{ x_{1:n} \in \mathcal{X}^n : 2^{-n(H_2(X)+\epsilon)} \le p(x_{1:n}) \le 2^{-n(H_2(X)-\epsilon)} \right\}.$$

In this set, all tuples have roughly the same probability mass. This is know as the *Asymptotic Equipartition property* (AEP). Here is an intuition of this typical set:

- For the low probability tuples $p(x_{1:n}) < 2^{-n(H_2(X)+\epsilon)}$, they are too unlikely to matter;
- For the high probability tuples $p(x_{1:n}) > 2^{-n(H_2(X)-\epsilon)}$, they are too few to matter;
- Therefore, we exclude those unimportant tuples and retain only the average probability tuples.

We now study the size of the reduced set.

Proposition 1.16. *Let $A_\epsilon^{(n)}$ be the ϵ-typical set for $X_{1:n}$. Then there exists $N_\epsilon > 0$ such that*

$$\mathbb{P}\left(X_{1:n} \in A_\epsilon^{(n)}\right) \ge 1 - \epsilon, \quad \forall n \ge N_\epsilon.$$

Furthermore, the upper bound of the typical set is given by

$$\left|A_\epsilon^{(n)}\right| \le 2^{n(H_2(X)+\epsilon)}, \quad \forall n \ge 1;$$

and the lower bound of the typical set is given by

$$\left|A_\epsilon^{(n)}\right| \ge (1-\epsilon) 2^{n(H_2(X)-\epsilon)}, \quad \forall n \ge N_\epsilon.$$

Proof. For the upper bound, note that

$$1 = \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n}) \ge \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) \ge \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X)+\epsilon)}.$$

For the lower bound, when $n \ge N_\epsilon$, we have

$$1 - \epsilon \le \mathbb{P}\left(X_{1:n} \in A_\epsilon^{(n)}\right) = \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) \le \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X)-\epsilon)}.$$

Rearranging each inequality completes the proof. □

1.4 Entropy Rates

In this section, we study a discrete-time stochastic process $X = (X_t)_{t\in\mathbb{N}}$, where each X_t is a random variable in a finite range $\mathcal{X}$. These random variables do not need to be i.i.d..

Definition 1.17. *Let $X = (X_t)_{t\in\mathbb{N}}$ be a stochastic process.*

(i) Average entropy per symbol

$$H(X) = \lim_{n\to\infty} \frac{H(X_{1:n})}{n}$$

(ii) The k-th order entropy

$$H^k(X) = H(X_k|X_{k-1}, \cdots, X_1)$$

(iii) Rate of information innovation

$$H^\infty(X) = \lim_{k\to\infty} H^k(X) = \lim_{k\to\infty} H(X_k|X_{k-1}, \cdots, X_1)$$

Remark. If $X = (X_t)_{t\in\mathbb{N}}$ is an i.i.d. sequence, we have

$$H(X) = H^\infty(X) = H(X_1).$$

Stationarity. Recall that a stochastic process $X = (X_t)_{t\in\mathbb{N}}$ is said to be *(strongly) stationary* if

$$\mathbb{P}\left(X_1 = x_1, \cdots, X_n = x_n\right) = \mathbb{P}\left(X_{k+1} = x_1, \cdots, X_{n+k} = x_n\right)$$

for every $n \in \mathbb{N}$, every lapse $k \in \mathbb{N}$ and all $x_1, \cdots, x_n \in \mathcal{X}$.

Theorem 1.18. *For a stationary process $X = (X_t)_{t\in\mathbb{N}}$,*

$$H(X) = H^\infty(X).$$

Proof. We first prove the existence of rate of information innovation. By stationarity,

$$H^n(X) = H(X_n|X_{n-1}, \cdots, X_2, X_1) \le H(X_n|X_{n-1}, \cdots, X_2) = H(X_{n-1}|X_{n-2}, \cdots, X_1)$$

Therefore, $H(X_n|X_{n-1}, \cdots, X_1)$ is decreasing in n. Since conditional entropy is nonnegative, the monotone sequence converges: $H^n \searrow H^\infty$. Next, by the chain rule of entropy,

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

The right-hand side of the last display, which is a Cesàro mean, has the same limit as $H(X_n|X_{n-1}, \cdots, X_1)$, which is $H^\infty(X)$. Since the limit of the left-hand side is the average entropy per symbol, the result follows. $\square$

Kolmogorov extension. If $(X_t)_{t\in\mathbb{N}}$ is a stationary process, then all finite-dimensional marginal distributions of this process are determined. By Kolmogorov extension theorem, we can extend the index of this process to the integer set $\mathbb{Z}$ and obtain a stationary process $(X_t)_{t\in\mathbb{Z}}$. We write for the past history

$$X_{\le 0} = (X_t)_{t\in -\mathbb{N}_0} = (X_0, X_{-1}, X_{-2}, \cdots).$$

Furthermore, we can define the conditional p.m.f. of X_1 given $X_{\leq 0}$:

$$\begin{aligned}
p(x_1|X_{\leq 0}) = \mathbb{E}\left[\mathbb{1}_{\{X_1=x_1\}}|X_{\leq 0}\right] &= \lim_{n\to\infty}\left[\mathbb{1}_{\{X_1=x_1\}}|X_0, X_{-1}, \cdots, X_{-n}\right] \\
&= \lim_{n\to\infty} p(x_1|X_0, X_{-1}, \cdots, X_{-n}).
\end{aligned}$$

Here the convergence holds both in L^1 and almost surely, since the sequence we take limit of is a uniformly integrable martingale. Furthermore, by Lebesgue's dominated convergence theorem,

$$\mathbb{E}\left[-\log p(X_1|X_{\leq 0})\right] = \lim_{n\to\infty} H^k(X) = H^\infty(X).$$

Ergodicity. Let $(\Omega, \mathscr{F}, P)$ be a measure space. A measurable mapping $T : (\Omega, \mathscr{F}) \to (\Omega, \mathscr{F})$ is said to be *ergodic*, if every set $A \in \mathscr{F}$ such that $TA = A$ *a.e.* satisfies $P(A) = 0$ or $P(A) = 1$. We let T play a role of time shift. The stochastic process $X = (X_t)_{t\in\mathbb{N}}$ is said to be an *ergodic* process, where $X_t(\omega) = X_0(T^t\omega)$ for all $t \in \mathbb{N}$ and $X_0 : \Omega \to \mathcal{X}$ is a random variable.

According to *Birkhoff's ergodic theorem*, the strong law of large numbers holds for a stationary ergodic process $X = (X_t)_{t\in\mathbb{N}}$:

$$\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k \to \mu = \mathbb{E}X_1, \quad a.s..$$

Lemma 1.19. *For the process $(X_t)_{t\in\mathbb{Z}}$, define the k-th order Markov approximation by*

$$p^k(X_{1:n}) = p(X_{1:k})\prod_{j=k+1}^{n} p(X_j|X_{j-1}, \cdots, X_{j-k}).$$

If $(X_t)_{t\in\mathbb{Z}}$ is a stationary ergodic process,

$$\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} \to H^k(X) \ a.s., \quad \textit{and} \quad \frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\leq 0})} \to H^\infty(X) \ a.s..$$

Proof. Since $(X_t)_{t\in\mathbb{Z}}$ is an ergodic process, so is the process $Y_t = f(X_{\leq t})$, where f is any measurable function. Then both $\log p(X_n|X_{n-1}, \cdots, X_{n-k})$ and $\log p(X_n|X_{\leq n-1})$ are stationary ergodic processes on $n \in \mathbb{N}$. By Birkhoff's ergodic theorem, we have

$$\begin{aligned}
&\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} = \frac{1}{n}\log\frac{1}{p(X_{1:k})} + \frac{1}{n}\sum_{j=k+1}^{n}\log\frac{1}{p(X_j|X_{j-1}, \cdots, X_{j-k})} \to 0 + H^k(X), \ a.s., \\
&\frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\leq 0})} = \frac{1}{n}\sum_{j=1}^{n}\log\frac{1}{p(X_j|X_{\leq j-1})} \to H^\infty(X), \ a.s..
\end{aligned}$$

Then we complete the proof. □

Lemma 1.20 (Sandwich)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process. Then*

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \leq 0 \ a.s., \quad \limsup_{n\to\infty}\frac{1}{n}\log\frac{p(X_{1:n})}{p(X_{1:n}|X_{\leq 0})} \leq 0 \ a.s..$$

Proof. Let A be the support set of $p(x_{1:n})$. Then

$$\mathbb{E}\left[\frac{p^k(X_{1:n})}{p(X_{1:n})}\right] = \sum_{x_{1:n}\in A}\frac{p^k(x_{1:n})}{p(x_{1:n})}p(x_{1:n}) = \sum_{x_{1:n}\in A} p^k(x_{1:n}) \leq \sum_{x_{1:n}\in\mathcal{X}^n} p^k(x_{1:n}) = 1.$$

By Markov's inequality, we have

$$\mathbb{P}\left(\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq \frac{2\log n}{n}\right) = \mathbb{P}\left(\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq n^2\right) \leq \frac{1}{n^2}$$

By Borel-Cantelli Lemma, since $\sum_{n=1}^{\infty} n^{-2} < \infty$, the events

$$\left\{\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq \frac{2\log n}{n}, \quad n \in \mathbb{N}\right\}$$

happens finitely many times with probability 1, which proves the first result. On the other hand, let $B(X_{\leq 0})$ be the support set of $p(x_{1:n}|X_{\leq 0})$. Then

$$\mathbb{E}\left[\frac{p(X_{1:n})}{p(X_{1:n}|X_{\leq 0})}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{p(X_{1:n})}{p(X_{1:n}|X_{\leq 0})}\middle| X_{\leq 0}\right]\right] = \mathbb{E}\left[\sum_{x_{1:n}\in B(X_{\leq 0})} p(X_{1:n})\right] \leq 1.$$

The second result then follows from a similar procedure. $\square$

Now we point out that, the Asymptotic Equilibrium property holds not only for i.i.d. sequences, but also for stationary ergodic processes.

Theorem 1.21 (Shannon-McMillan-Breiman)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process. Then*

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} = H^{\infty}(X).$$

Proof. By Lemmas 1.19 and 1.20, almost surely,

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} &\leq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} = H^k(X),\\ \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} &\geq \limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\leq 0})} = H^{\infty}(X).\end{aligned}$$

Therefore, for all $k \in \mathbb{N}$, we have

$$H^{\infty}(X) \leq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \leq \limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \leq H^k(X).$$

Since X is stationary, $H^k(X) \searrow H^{\infty}(X)$ as $k \to \infty$. Hence $\frac{1}{n}\log\frac{1}{p(X_{1:n})} \overset{a.s.}{\to} H^{\infty}(X)$. $\square$

Remark. An example for stationary ergodic process is the irreducible and aperiodic Markov chain.

2 Lossless Compression

In this section, we study the problem of lossless coding. To begin with, we have a source alphabet $\mathcal{X}$ and a D-ary alphabet $\{0, 1, \cdots, D-1\}$. Our key goal is to transform a string of $\mathcal{X}$ to a string of $\mathcal{D}$.

- A *source code* is a mapping $C : \mathcal{X} \to \mathcal{D}^*$, where $\mathcal{D}$ is a D-ary alphabet $\{0, 1, \cdots, D-1\}$, and

$$\mathcal{D}^* = \bigcup_{n=1}^{\infty} \mathcal{D}^n.$$

  The elements of $C(\mathcal{X})$ are called *codewords*. For every symbol $x \in \mathcal{X}$, we denote by $\ell(x)$ the length of the codeword $C(x)$ associated with x.
- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *nonsingular* if it is injective.
- The *extension* $C^* : \mathcal{X}^* \to \mathcal{D}^*$ of a source code C is the mapping from finite length strings of $\mathcal{X}$ to finite length strings of $\mathcal{D}$:

$$C^*(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n).$$

- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *uniquely decodable* if its extension C^* is injective.
- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *instantaneous* (or *prefix-free*) if no codeword of C is prefixed by any other codeword.
- We have the inclusions: *nonsingular codes* $\supset$ *uniquely decodable codes* $\supset$ *instantaneous codes*.

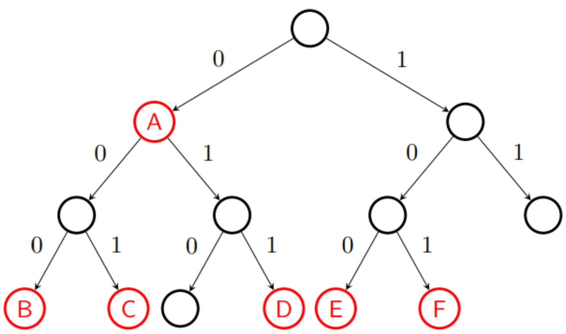
In general, some nice properties of a code are wanted:

- it is uniquely decodable;
- it is prefix free, so one can decode a string instantaneously while reading;
- it is efficient, i.e. given the distribution p of letters $\mathcal{X}$ in a string, we would like to minimize the average codeword length:

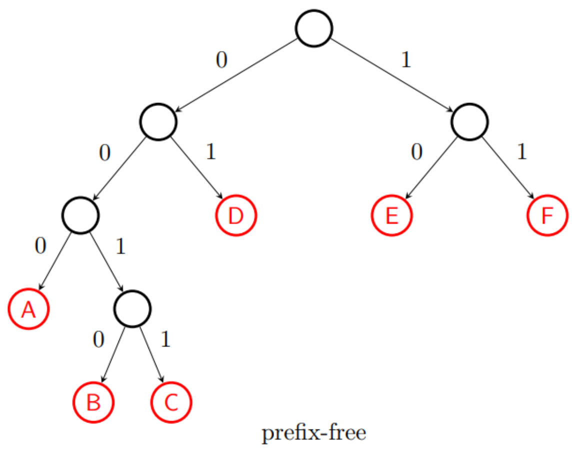
$$\mathbb{E}\left[\ell(X)\right] = \sum_{x \in \mathcal{X}} p(x)\ell(x).$$

2.1 Kraft-McMillan Inequality

Tree representation. A D-ary code $C : \mathcal{X} \to \mathcal{D}$ can be represented as a D-ary tree that consists of a root with branches, nodes and leaves. The root and every node has exactly D children, with each branch labeled by a letter in $\mathcal{D}$. Starting from the root, each vertex is uniquely associated with a string $d \in \mathcal{D}^*$, specified by the path from the root to itself. Some examples of binary trees are given below.



nonsingular but not prefix-free



prefix-free

We can determine whether a code is instantaneous right away by looking at its tree.

Proposition 2.1. *A code $C : \mathcal{X} \to \mathcal{D}^*$ is instantaneous if and only if all its codeword are leaves.*

Proof. If $C : \mathcal{X} \to \mathcal{D}^*$ is an instantaneous code, then each of its codeword has no descendant in the tree, which is a leaf; conversely, if each codeword of C is a leaf in the tree, it has no ancestor which is also a codeword, and C is instantaneous. $\square$

Using the tree representation, we can show a property which characterizes the instantaneous codes.

Theorem 2.2 (Kraft's inequality)**.** *Let $\ell : \mathcal{X} \to \mathbb{N}$ be a length function. Then ℓ is the length function of an instantaneous code if and only if it satisfies Kraft's inequality:*

$$\sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1. \tag{2.1}$$

Proof. We first prove necessity. Let ℓ is the length function of an instantaneous code C, and let L be the depth of the tree. Then every codeword $C(x)$ at depth $\ell(x)$ prunes away $D^{L-\ell(x)}$ leaves from the complete tree of depth L. Since there are no more than D^L leaves in the complete tree, we have

$$\sum_{x\in\mathcal{X}} D^{L-\ell(x)} \le D^L \quad \Rightarrow \quad \sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1.$$

Now we prove the sufficiency. To this end, we prove the following argument: at every step $k \in \mathbb{N}$, after all codewords of length $\ell(x) < k$ have been assigned, there is enough room left at the depth k for the codewords of length $\ell(x) = k$. More explicitly, we want to show

$$D^k - \sum_{x\in\mathcal{X}:\ell(x)<k} D^{k-\ell(x)} \ge \left|C^{-1}(\mathcal{D}^k)\right|, \quad \forall 1 \le k \le L.$$

Note that

$$\left|C^{-1}(\mathcal{D}^k)\right| = \sum_{x\in\mathcal{X}:\ell(x)=k} D^{k-\ell(x)}.$$

Then our conclusion holds if

$$\sum_{x\in\mathcal{X}:\ell(x)\le k} D^{-\ell(x)} \le 1, \quad \forall k \in \mathbb{N}.$$

Clearly this is valid by Kraft's inequality (2.1). $\square$

The Kraft's inequality is also a necessary condition for a code to be uniquely decodable.

Theorem 2.3 (McMillan)**.** *Every uniquely decodable code $C : \mathcal{X} \to \mathcal{D}^*$ satisfies Kraft's inequality (2.1).*

Proof. Let $C : \mathcal{X} \to \mathcal{D}^*$ be a uniquely decodable code, and let $L = \max_{x\in\mathcal{X}} \ell(x)$, where ℓ is the length function of C. Then for a source string $x_{1:n}$, the length of the extended codeword $C^*(x_{1:n})$ is given by

$$\ell^*(x_{1:n}) = \sum_{i=1}^{n} \ell(x_i) \le nL.$$

Let N_k be the number of source strings of length n with $\ell^*(x_{1:n}) = k$. Since C is uniquely decodable, the source strings with codewords of length k are no more than D-ary strings of length k, i.e. $N_k \le D^k$. Then

$$\sum_{x_{1:n}\in\mathcal{X}^n} D^{-\ell^*(x_{1:n})} = \sum_{k=1}^{nL} N_k D^{-k} \le \sum_{k=1}^{nL} D^k D^{-k} \le nL.$$

On the other hand,

$$\begin{aligned}\sum_{x_{1:n}\in\mathcal{X}^n} D^{-\ell^*(x_{1:n})} &= \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_n\in\mathcal{X}} D^{-\ell(x_1)}D^{-\ell(x_2)}\cdots D^{-\ell(x_n)} \\ &= \sum_{x_1\in\mathcal{X}} D^{-\ell(x_1)}\sum_{x_2\in\mathcal{X}} D^{-\ell(x_2)}\cdots\sum_{x_n\in\mathcal{X}} D^{-\ell(x_n)} = \left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^n.\end{aligned}$$

Therefore, we have

$$\sum_{x\in\mathcal{X}} D^{-\ell(x)} \le \inf_{n\in\mathbb{N}} \sqrt[n]{nL} = 1.$$

Then we complete the proof. □

Remark. To summarize, the Kraft's inequality (2.1) is a

- sufficient condition for the existence of an instantaneous code;
- necessary condition for a code to be uniquelt decodable.

2.2 Fundamental Limits of Compression

In this section, we study the limits of lossless compression. Given a source distribution p on $\mathcal{X}$, we want to minimize the average codeword length of our code. By Kraft-McMillan inequality, the search for optimal code can be expressed as the following optimization problem:

$$\min_{l:\mathcal{X}\to\mathbb{N}} \sum_{x\in\mathcal{X}} p(x)\ell(x) \quad \textit{subject to} \quad \sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1.$$

Following is a fundamental result of lossless compression.

Theorem 2.4. *For any source distribution $X\sim p$ on $\mathcal{X}$, the expected length $\mathbb{E}[\ell(X)]$ of an optimal uniquely decodable D-ary code satisfies*

$$\frac{H(X)}{\log D} \le \mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 1. \tag{2.2}$$

Proof. Upper bound. By Theorem 2.2, it suffices to construct a length function $\ell:\mathcal{X}\to\mathbb{N}$ that satisfies both the Kraft's inequality and the second (strict) inequality given in (2.2). Consider Shannon's length function:

$$\ell(x) = \left\lceil \log_D \frac{1}{p(x)} \right\rceil, \quad x\in\mathcal{X}, \tag{2.3}$$

Since

$$\sum_{x\in\mathcal{X}} D^{-\ell(x)} \le \sum_{x\in\mathcal{X}} D^{\log_D p(x)} = \sum_{x\in\mathcal{X}} p(x) = 1,$$

there exists an instantaneous code $C:\mathcal{X}\to\mathcal{D}^*$ whose length function is ℓ. On the other hand,

$$\mathbb{E}[\ell(X)] = \sum_{x\in\mathcal{X}} p(x)\ell(x) < \sum_{x\in\mathcal{X}} p(x)\left(\log_D\frac{1}{p(x)} + 1\right) = \frac{H(X)}{\log D} + 1.$$

Hence the upper bound holds.

Lower bound. We consider the following relaxed optimization problem:

$$\min_{l:\mathcal{X}\to\mathbb{R}} \sum_{x\in\mathcal{X}} p(x)\ell(x) \quad \textit{subject to} \quad \sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1.$$

Note that the range of ℓ is $\mathbb{R}_+$. The Lagrange function is

$$L(l,\lambda) = \sum_{x\in\mathcal{X}} p(x)\ell(x) + \lambda\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1\right),$$

with KKT conditions

$$\begin{cases} \frac{\partial L}{\partial l(x)} = p(x) - \lambda D^{-l(x)}\log D = 0, \\ \lambda \ge 0,\ \sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1 \le 0, \\ \lambda\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1\right) = 0. \end{cases}$$

The optimal solution is given by

$$\lambda = \frac{1}{\log D}, \quad l(x) = \log_D \frac{\lambda \log D}{p(x)} = \log_D \frac{1}{p(x)},\ x\in\mathcal{X},$$

and the optimal value is

$$\sum_{x\in\mathcal{X}} p(x)\ell(x) = \sum_{x\in\mathcal{X}} p(x)\log_D \frac{1}{p(x)} = \frac{H(X)}{\log D}. \tag{2.4}$$

Since our problem is relaxed, the primal problem (2.3) has optimal value no less than (2.4). Hence the lower bound holds for all uniquely decodable codes. □

Remark. In fact, we proved the existence of an *instantaneous* code with

$$\mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 1.$$

Coding over blocks. Using integer codeword lengths may lead to waste of memory. To overcome this effect, we consider coding over blocks of input symbols. If the input data $X_1, X_2, \cdots$ is an i.i.d. sequence of symbols, we partition it into blocks of size n and create a new source $\widetilde{X}_1, \widetilde{X}_2, \cdots$, where

$$\widetilde{X}_1 = (X_1,\cdots,X_n),\ \widetilde{X}_2 = (X_{n+1},\cdots,X_{2n}),\ \cdots,\ \widetilde{X}_k = (X_{(k-1)n+1},\cdots,X_{kn}),\ \cdots.$$

Consequently, every vector $\widetilde{X}_k$ can be viewed as a symbol from the alphabet $\widetilde{\mathcal{X}} = \mathcal{X}^n$, and we can find an optimal code $\widetilde{C} : \widetilde{X} \to \mathcal{D}$, whose length function ℓ satisfies

$$\frac{H(\widetilde{X})}{\log D} \le \mathbb{E}\left[\ell(\widetilde{X})\right] \le \frac{H(\widetilde{X})}{\log D} + 1.$$

Note that $H(\widetilde{X}) = nH(X)$, the average codeword length per symbol (in $\mathcal{X}$) satisfies

$$\frac{H(X)}{\log D} \le \frac{1}{n}\mathbb{E}\left[\ell(\widetilde{X})\right] < \frac{H(X)}{\log D} + \frac{1}{n}.$$

As the block size n increases, the integer effect becomes negligible. However, we also introduce delay in our system and increase the complexity of our code.

2.3 Shannon-Fano-Elias Coding

In this section, we introduce a specific coding approach that is near-optimal.

Midpoints of CDF. Without loss of generality, we assume that the source alphabet is $\mathcal{X} = \{1, 2, \cdots, m\}$, and $p(1) \geq p(2) \geq \cdots \geq p(m)$. The cumulative distribution function of p is

$$F(r) = \sum_{j=1}^{m} \mathbb{1}_{\{j \leq r\}} p(j), \quad r \in \mathbb{R}.$$

We define $\overline{F}(x)$ to be the midpoint of the interval $[F(x-1), F(x))$:

$$\overline{F}(x) = \sum_{j=1}^{x-1} p(j) + \frac{p(x)}{2}, \quad x = 1, \cdots, m.$$

Then $\overline{F}(x)$ is a real number in $(0, 1)$ that uniquely identifies $x \in \mathcal{X}$.

***D*-ary expansion and truncation.** The D-ary expansion of a real number $\overline{F}(x) \in (0, 1)$ is given by

$$\overline{F}(x) = (0.z_1 z_2 \cdots)_D = \sum_{k=1}^{\infty} z_k D^{-k} = z_1 D^{-1} + z_2 D^{-2} + \cdots, \quad z_1, z_2, \cdots \in \{0, 1, \cdots, D-1\}.$$

Given a positive integer $\ell \in \mathbb{N}$, one have the ℓ-truncation of the D-ary expansion of $\overline{F}(x)$:

$$C(x) = (0.z_1 z_2 \cdots z_\ell)_D = \sum_{k=1}^{\ell} z_k D^{-k}$$

To ensure that the codeword of x is unique, we let $\overline{F}(x) - C(x) < \frac{p(x)}{2}$, so that

$$C(x-1) \leq \overline{F}(x-1) < F(x-1) < C(x).$$

To this end, we set

$$\ell = \left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1,$$

then

$$\overline{F}(x) - C(x) < D^{-\ell} \leq D^{-\log_D \frac{1}{p(x)} - 1} \leq \frac{p(x)}{D} \leq \frac{p(x)}{2}.$$

Construction of the Shannon-Fano-Elias code. For each $x \in \mathcal{X}$:

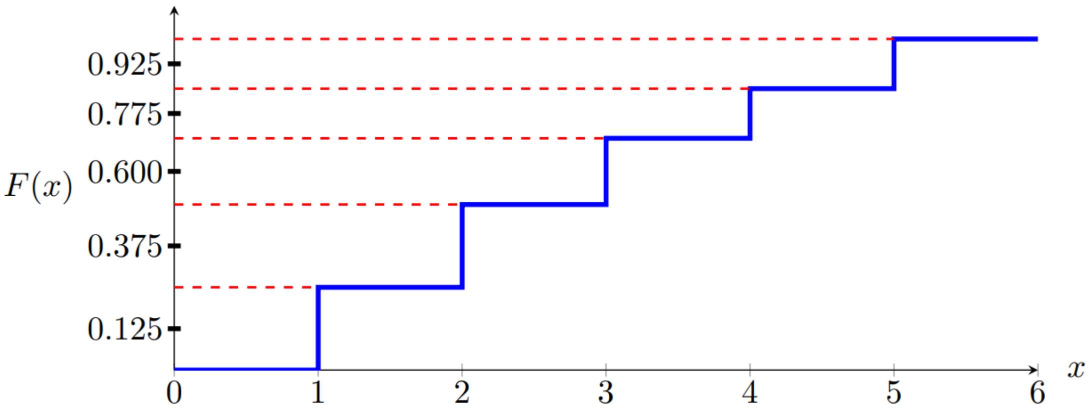
- Let z be the D-ary expansion of x;
- Choose the length of the codeword of x:

$$\ell(x) = \left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1;$$

- Choose the codeword of x to be the first most significant D-ary digits:

$$z = 0.\underbrace{z_1 z_2 \cdots z_{\ell(x)}}_{C(x)} z_{\ell(x)+1} \cdots.$$

An example of binary Shannon code. Here we let $\mathcal{X} = \{1, 2, 3, 4, 5\}$, and $D = 2$.



x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ in binary	$\ell(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$	codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.25	0.5	0.375	0.011	3	011
3	0.2	0.7	0.6	0.10011	4	1001
4	0.15	0.85	0.775	0.1100011	4	1100
5	0.15	1.0	0.925	0.1110110	4	1110

Shannon code is instantaneous. If the codeword $C(x) = (0.z_1 \cdots z_{\ell(x)})_D$ is a prefix of another codeword, this codeword lies in the half-open interval

$$\left[(0.z_1 \cdots z_{\ell(x)})_D, (0.z_1 \cdots z_{\ell(x)})_D + \frac{1}{D^{l(x)}}\right).$$

However, a contradiction rises because

$$C(x+1) - C(x) > F(x) - \overline{F}(x) = \frac{p(x)}{2} \geq D^{-l(x)}.$$

Average codeword length. The average codeword length is given by

$$\mathbb{E}[\ell(X)] = \sum_{x \in \mathcal{X}} p(x) \left(\left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1 \right),$$

which satisfies

$$\frac{H(X)}{\log D} + 1 \leq \mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 2.$$

It is revealed that the Shannon code is sub-optimal.

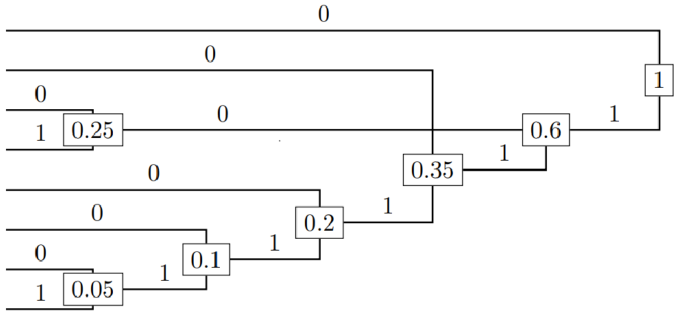
2.4 Huffman Coding

The search for binary optimal code was discovered by David Huffman (1952).

Construction of Huffman tree. The construction procedure is greedy.

- Take the two least probable symbols, which will be assigned the longest codewords having equal lengths and differing only at the last digit;
- Merge these two symbols into a new symbol with combined probability mass and repeat.

Codeword	x	$p(x)$
0	1	0.40
110	2	0.15
100	3	0.15
101	4	0.10
1110	5	0.10
11110	6	0.05
111110	7	0.04
111111	8	0.01



Optimality of Huffman code. Let $\mathcal{X} = \{1, 2, \cdots, m\}$. Without loss of generality, assume probabilities are in descending order $p(1) \geq p(2) \geq \cdots \geq p(m)$. We prove the optimality of Huffman code through three step.

Lemma 2.5. *In an optimal code, shorter codewords are assigned larger probabilities, i.e. $p(i) > p(j)$ implies $\ell(i) \leq \ell(j)$.*

Proof. Argue by contradiction. If there exists $i, j \in \mathcal{X}$ with $\ell(i) \leq \ell(j)$ and $p(i) > p(j)$, then we can exchange these codewords and reduce the expected length. Hence the code is not optimal. $\square$

Lemma 2.6. *There exists an optimal code for which the codewords assigned to the smallest probabilities are siblings, i.e., they have the same length and differ only in the last symbol.*

Proof. Consider any optimal code. By Lemma 2.5, the codeword $C(m)$ has the longest length. Assume for the sake of contradiction, its sibling is not a codeword. Then the expected length can be decreased by moving $C(m)$ to its parent. Thus, the code is not optimal and a contradiction is reached.

Now, we know the sibling of $C(m)$ is a codeword. If it is $C(m-1)$, we are done. If it is some $C(i)$ for $i \neq m-1$ and the code is optimal, by Lemma 2.5, we have $p(i) = p(m-1)$. Therefore, $C(i)$ and $C(m-1)$ can be exchanged without changing expected length. $\square$

Theorem 2.7 (Optimality of Huffman coding)**.** *Huffman's coding algorithm produces an optimal code tree.*

Proof. Let ℓ be the length function of the optimal code. By Lemmas 2.5 and 2.6, $C(m)$ and $C(m-1)$ are siblings and the longest codewords. Then we merge the two symbols and let $\widetilde{p}_1 \geq \cdots \geq \widetilde{p}_{m-1}$ denote the reordered probabilities after merging $p(m)$ and $p(m-1)$, and denote by $\widetilde{C}_1, \cdots, \widetilde{C}_{m-1}$ the corresponding codewords. The reduced length function $\widetilde{\ell}$ satisfies

$$\mathbb{E}\left[\ell(X)\right] = \mathbb{E}\left[\widetilde{\ell}(\widetilde{X})\right] + \mathbb{P}\left(\ell(X) \neq \widetilde{\ell}(\widetilde{X})\right) = \mathbb{E}\left[\widetilde{\ell}(\widetilde{X})\right] + p(m-1) + p(m).$$

Hence ℓ is the length function of an optimal code if and only if $\widetilde{\ell}$ is the length function of an optimal code for the reduced alphabet. The problem then is reduced to finding an optimal code tree for $\widetilde{p}_1 \geq \cdots \geq \widetilde{p}_{m-1}$. Repeat the merging procedure above for m times, and the result follows. $\square$