Fantuan's Academia

FANTUAN'S SCRIBED NOTES SERIES VOL.2

Scribed Notes on Information Theory

Author: Jingxuan Xu
Lecturer: Omar Montasser

April 25, 2025

Contents

All the Sections with * are hard sections and can be skipped without losing coherence.

This note is referenced on **Elements of Information Theory**[1] by Cover and Thomas, **Information Theory: From Coding to Learning**[2] by Polyanskiy and Wu, and S&DS 664 Information Theory course taught by Prof. Omar Montasser at Yale University.

Chapter 1

Information Measures

1.1 Why Information Theory?

What is Information?

By L. Brillouin: "*We must start with a precise definition of the word "information". We consider a problem involving a certain number of possible answers, if we have no special information on the actual situation. When we happen to be in possession of some information on the problem, the number of possible answers is reduced, and complete information may even leave us with only one possible answer.*" Therefore, we define **information** as **a measure of difference between two beliefs about the system state.**

What is Information Theory?

In the narrowest sense, it is a scientific discipline concerned with *optimal methods of transmitting and storing data.* The highlights of this part of the subject are so called "coding theorems" showing existence of algorithms for compressing and communicating information across noisy channels. However, information theory is not limited to the data compression and transmission tasks, because the true scope of the field is much broader. It has broad application in fields such as probability theory, statistics, physics, computer science and economics.

1.2 Entropy

Entropy is a **measure of uncertainty** of a random variable.

Definition 1.2.1: Entropy

Let X be a discrete random variable with probability mass function $p_X(x) = \mathbb{P}[X = x]$, $x \in \mathcal{X}$. The **(Shannon) entropy** $H(X)$ of r.v. X is defined as

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

Here are few things to note:

- When computing the sum, by continuity of $x \mapsto x \log \frac{1}{x}$, we agree that $0 \log 0 = 0$ since $x \log x \to 0$ as $x \to 0$. Therefore, adding terms of zero probability does not change the entropy.
- Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities. Therefore, we also write $H(p_X)$, or simply $H(p)$ for the above quantity.
- The basis of the logarithm determines the unit of the entropy. $\log_2$ is in *bits*, $\log_e$ is in *nats*, $\log_{256}$ is in *bytes*. We commonly denote entropy with base b by $H_b(X)$.

Few results follow immediately.

Lemma 1.2.2: Positivity of Entropy

$H(X) \geqslant 0$.

Proof. $p_X(x) \geqslant 0$. Also, $0 \leqslant p_X(x) \leqslant 1$, so $\log p_X(x) \leqslant 0$. Thus, $H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \geqslant 0$. □

Lemma 1.2.3: Change of Basis

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

Proof. By the fact that $\log_b p = \log_b a \log_a p$. □

Example 1: Let X be Bernoulli distributed with parameter p, i.e., $X \sim \text{Ber}(p)$. Then,

$$H(X) = -p \log p - (1-p) \log(1-p) := h(p)$$

This function is important and will be used several times, so we give it a notation $h(p)$. Below is the plot of the entropy w.r.t. p with unit in bits (i.e., $\log_2$ base). This makes sense because, when $p = 0$ or 1, the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximized when $p = 1/2$ intuitively.

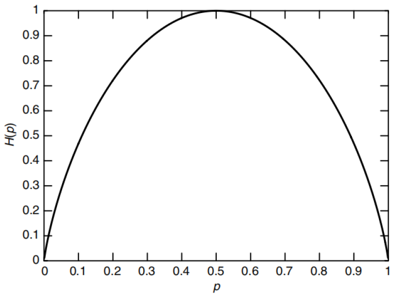


Figure 1.1: Entropy of Bernoulli Random Variable

Example 2: Let

$$X = \begin{cases} a, & \text{with probability } 1/2 \\ b, & \text{with probability } 1/4 \\ c, & \text{with probability } 1/8 \\ d, & \text{with probability } 1/8 \end{cases}$$

The entropy of X with base 2 is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - 2\frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits}$$

Example 3: Let X be geometrically distributed with parameter p, i.e., $X \sim \text{Geom}(p)$, $p_X(x) = p(1-p)^x$, $x = 0, 1, \cdots$. Then, $\mathbb{E}[X] = (1-p)/p$, and

$$H(X) = \mathbb{E}\left[\log\left(\frac{1}{p(1-p)^X}\right)\right] = \mathbb{E}\left[\log\frac{1}{p}\right] + \mathbb{E}\left[X\log\frac{1}{1-p}\right] = \log\frac{1}{p} + \mathbb{E}[X]\log\frac{1}{1-p} = \frac{h(p)}{p}$$

Example 4: It is also possible that $H(X) = +\infty$. For example, when $\mathbb{P}[X = k] \propto \frac{1}{k\log^2 k}$, $k = 2, 3, \cdots$, we have

$$\begin{aligned} H(X) &= -\sum_{k=2}^{\infty}\mathbb{P}[X=k]\log\mathbb{P}[X=k] = -\sum_{k=2}^{\infty}\frac{1}{C}\frac{1}{k\log^2 k}\log\left(\frac{1}{C}\frac{1}{k\log^2 k}\right) \\ &= -\sum_{k=2}^{\infty}\frac{1}{C}\frac{1}{k\log^2 k}\left(-\log C - \log k - 2\log(\log k)\right) \\ &= \underbrace{\frac{\log C}{C}\sum_{k=2}^{\infty}\frac{1}{k\log^2 k}}_{\text{Term 1}} + \underbrace{\frac{1}{C}\sum_{k=2}^{\infty}\frac{1}{k\log k}}_{\text{Term 2}} + \underbrace{\frac{2}{C}\sum_{k=2}^{\infty}\frac{\log(\log k)}{k\log^2 k}}_{\text{Term 3}} \end{aligned}$$

Since $1/k\log k$ is convex, by integration test, $\sum_{k=2}^{\infty}\frac{1}{k\log k} > \int_2^{\infty}\frac{1}{x\log x}\,\mathrm{d}x = \lim_{t\to\infty}\log(\log x)|_2^t = \infty$, the Term 2 diverges, so the whole formula diverges. (Graphically, see BV1z54y1X7Sj)

Remark: To understand the operational meaning of entropy, let us consider the following game. We are allowed to *make queries about some unknown discrete r.v. X by asking yes-no questions.* The objective of the game is to guess the realized value of the r.v. X. For example, in Example 2, we can ask '$X = a$?'. If not, proceed by asking '$X = b$?'. If not, ask '$X = c$?', after which we will know for sure the realization of X. The resulting average number of questions is $1/2 + 1/4 \times 2 + 1/8 \times 3 = 1.75$, which equals $H(X)$ in bits. We will show later that the minimal average number of yes-no questions to pin down the value of X is always between $H(X)$ bits and $H(X) + 1$ bits.

1.3 Joint and Conditional Entropy

We extend the definition of entropy to multiple random variables. There is nothing really new in this definition because this below ((X, Y), and $(X_1, \cdots, X_n)$) can be considered to be a single vector-valued random variable.

Definition 1.3.1: Joint Entropy

- (pair) The **joint entropy** $H(X, Y)$ of a pair of discrete r.v. (X, Y) with joint distribution $p_{X,Y}(x, y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p_{X,Y}(x,y)\log p_{X,Y}(x,y) = -\mathbb{E}\left[\log p_{X,Y}(X,Y)\right]$$

- (multi) The **joint entropy** of n discrete r.v.s $(X_1, \cdots, X_n)$ with joint distribution $p_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$

is defined as

$$\begin{aligned} H(X_1,\cdots,X_n) &= -\sum_{x_1}\cdots\sum_{x_n} p_{X_1,\cdots,X_n}(x_1,\cdots,x_n)\log p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) \\ &= -\mathbb{E}\left[\log p_{X_1,\cdots,X_n}(X_1,\cdots,X_n)\right] \end{aligned}$$

We further define the conditional counterparts of entropy, by applying the original definition to a conditional probability measure followed by a further averaging. Conditional entropy measures the remaining randomness of a random variable when another is revealed. When Y depends on X, observing Y lower the entropy of X.

Definition 1.3.2: Conditional Entropy

If $(X, Y) \sim p_{X,Y}(x, y)$, denote $p_{Y|X}(y|x)$ the conditional distribution of Y given $X = x$, the **conditional entropy** $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) = \mathbb{E}_X\left[H(Y|X=x)\right] &= \sum_{x\in\mathcal{X}} p_X(x)H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p_X(x)\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x)\log p_{Y|X}(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log p_{Y|X}(y|x) \\ &= -\mathbb{E}_{X,Y}\left[\log p_{Y|X}(y|x)\right] \end{aligned}$$

Entropy follows the additive chain rule below:

Theorem 1.3.3: Chain Rule for Entropy

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

Proof. For simplicity, we omit the subscript of probability mass function here and after.

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

which shows the original equation. □

This can be generalized to n cases.

Theorem 1.3.4: Chain Rule for Entropy (Generalized)

Let $X_1, X_2, \cdots, X_n$ be r.v.s with joint pmf $p(x_1, x_2, \cdots, x_n)$. Then,

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

We use the convention that $H(X_i|X_{i-1}, \cdots, X_1) = H(X_1)$ when $i = 1$.

Proof. We can induct on the 2-dim case (which should be trivial), or we can use brute force:

$$\begin{aligned}
H(X_1, X_2, \cdots, X_n) &= -\sum_{x_1, x_2, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \log p(x_1, x_2, \cdots, x_n) \\
&= -\sum_{x_1, x_2, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \log \prod_{i=1}^{n} p(x_i|x_{i-1}, \cdots, x_1) && \text{(Bayes Rule)} \\
&= -\sum_{x_1, x_2, \cdots, x_n} \sum_{i=1}^{n} p(x_1, x_2, \cdots, x_n) \log p(x_i|x_{i-1}, \cdots, x_1) \\
&= -\sum_{i=1}^{n} \sum_{x_1, x_2, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \log p(x_i|x_{i-1}, \cdots, x_1) && \text{(Finite Sum Change Order)} \\
&= -\sum_{i=1}^{n} \sum_{x_1, x_2, \cdots, x_i} p(x_1, x_2, \cdots, x_i) \log p(x_i|x_{i-1}, \cdots, x_1) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)
\end{aligned}$$

where in the last row first equality we use the fact that fix i, $\log p(x_i|x_{i-1}, \cdots, x_1)$ is a constant, so we can sum up the joint pmf to x_i terms. □

Example 1: Let (X, Y) has the joint distribution

$$(X, Y) = \begin{cases} (0, 1), & \text{with probability } 1/3 \\ (1, 0), & \text{with probability } 1/3 \\ (1, 1), & \text{with probability } 1/3 \end{cases}$$

Then, the joint entropy of (X, Y), is

$$H(X, Y) = -3 \times \frac{1}{3} \log \frac{1}{3} = \log_2 3 \text{ bits}$$

Conditional on $X = 0$, Y is deterministic, so $H(Y|X = 0) = 0$. Conditional on $X = 1$, Y is Bernoulli with success probability $p = 1/2$. Therefore, as we have shown above, $H(Y|X = 1) = 1$ bits. Therefore, the conditional entropy of Y given X is

$$H(Y|X) = \mathbb{P}[X = 0]H(Y|X = 0) + \mathbb{P}[X = 1]H(Y|X = 1) = 0 + \frac{2}{3} \times 1 = \frac{2}{3} \text{ bits}$$

Example 2: Binary Noisy Channel Let $X \sim \text{Ber}(1/2)$, $Z \sim \text{Ber}(\delta)$ independent of X, where $\delta \in [0, 1]$. Let $Y = X \oplus Z$, where $\oplus$ denotes the binary addition function `XOR`. The chart below shows the relation between X, Y and Z.

X	Z	$Y = X \oplus Z$
1	1	0
0	1	1
1	0	1
0	0	0

We can see from the table that, when $Z = 1$ (first two rows, with probability δ), Y disagrees with X; when $Z = 0$ (last two rows, with proability $1 - \delta$), Y agrees with X. That's why it can be seen as a noisy channel with noise Z. We can get:

$$\mathbb{P}[X = 0|Y = 0] = 1 - \delta \text{ (probability of } Z = 0)$$

$$\mathbb{P}[X = 1|Y = 0] = \delta \text{ (probability of } Z = 1)$$

$$\mathbb{P}[X = 0|Y = 1] = \delta \text{ (probability of } Z = 1)$$

$$\mathbb{P}[X = 1|Y = 1] = 1 - \delta \text{ (probability of } Z = 0)$$

That is, $\mathbb{P}[X|Y = 0] = \text{Ber}(\delta)$ and $\mathbb{P}[X|Y = 1] = \text{Ber}(1 - \delta)$. Since $h(\delta) = h(1 - \delta)$, we have $H(X|Y = 0) = H(X|Y = 1) = h(\delta)$. We have shown that $H(X) = 1$ bit. Moreover, we note that

$$\mathbb{P}[Y = 0] = \mathbb{P}[X = 1, Z = 1] + \mathbb{P}[X = 0, Z = 0] = \mathbb{P}[X = 1]\mathbb{P}[Z = 1] + \mathbb{P}[X = 0]\mathbb{P}[Z = 0] = \frac{\delta}{2} + \frac{1 - \delta}{2} = \frac{1}{2}$$

$$\mathbb{P}[Y = 0] = \mathbb{P}[X = 1, Z = 0] + \mathbb{P}[X = 0, Z = 1] = \mathbb{P}[X = 1]\mathbb{P}[Z = 0] + \mathbb{P}[X = 0]\mathbb{P}[Z = 1] = \frac{1 - \delta}{2} + \frac{\delta}{2} = \frac{1}{2}$$

Then, the conditional entropy of X given Y can be calculated by

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) = \mathbb{P}[Y = 0]H(X|Y = 0) + \mathbb{P}[Y = 1]H(X|Y = 1) \\ &= \frac{1}{2}H(X|Y = 0) + \frac{1}{2}H(X|Y = 1) = \frac{1}{2}h(\delta) + \frac{1}{2}h(1 - \delta) = h(\delta) \end{aligned}$$

Note that when $\delta = 1/2$, Y is independent of X and $H(X|Y) = H(X) = 1$ bits. When $\delta = 0$ or 1, X is completely determined by Y and hence $H(X|Y) = 0$.

1.4 Relative Entropy (Kullback-Leibler Divergence)

Kullback-Leibler divergence is a distance measure between two distributions p and q. Though it is not rigorously a 'distance' (not symmetric, does not satisfy triangular inequality), it is still useful in many cases. In statistics, it arises as an expected logarithm of the likelihood ratio.

Definition 1.4.1: Relative Entropy (Kullback-Leibler Divergence)

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

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

We use the convention that $0 \log \frac{0}{0} = 0$ (if both have zero probability on a spot, there would be no contribution), $0 \log \frac{0}{q} = 0$, and $p \log \frac{p}{0} = \infty$ (these are based on continuity).

Remark: Due to the definition, the KL divergence can be infinity. For example, $\mathcal{X} = \{0, 1\}$, $p(0) = p(1) = 1/2$, $q(0) = 1$. Then, on $x = 1$, $p \neq 0$ but $q = 0$, which leads to infinity. Therefore, to make $D(p\|q)$ finite, we must have

$$\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$$

Example: Let $\mathcal{X} = \{0, 1\}$ and p, q are distributions on $\mathcal{X}$. Let $p(0) = 1 - r$, $p(1) = r$, and let $q(0) = 1 - s$, $q(1) = s$. Then,

$$D(p\|q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

and

$$D(q\|p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}$$

We define the conditional version of relative entropy.

Definition 1.4.2: Conditional Relative Entropy

For joint pmfs $p(x, y)$ and $q(x, y)$, the **conditional relative entropy** is the average of the relative entropies between the conditional pmfs $p(y|x)$ and $q(y|x)$ averaged over $p(x)$,

$$\begin{aligned} D\left(p(y|x)\|q(y|x)\right) &= \sum_{x_0 \in \mathcal{X}} p(x_0) D(p(y|x = x_0)\|q(y|x = x_0)) \\ &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log\left(\frac{p(y|x)}{q(y|x)}\right) = \mathbb{E}_{p(x,y)}\left[\log\left(\frac{p(Y|X)}{q(Y|X)}\right)\right] \end{aligned}$$

The corresponding chain rule for relative entropy is below.

Theorem 1.4.3: Chain Rule for Relative Entropy

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

Proof.

$$\begin{aligned} D(p(x, y)\|q(x, y)) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{q(x, y)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} && \text{(Bayes Rule)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x)}{q(x)} + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)) \end{aligned}$$

which shows the original equation □

1.5 Mutual Information

We now introduce mutual information, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition 1.5.1: Mutual Information

Let X, Y be two random variables with joint pmf $p(x,y)$ and marginal pmfs $p(x)$ and $p(y)$. Then, the **mutual information** $I(X;Y)$ is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)\|p(x)p(y)) = \mathbb{E}_{p(x,y)}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right]$$

The relation between entropy and mutual information is established below.

Proposition 1.5.2: Relation between Entropy and Mutual Information

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

(b) $I(X;X) = H(X)$.

Proof. (a)

$$\begin{aligned} I(X;Y) &\overset{\text{Bayes}}{=\!=} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(y)p(x|y)}{p(x)p(y)}\right) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x|y)}{p(x)}\right) \\ &= \underbrace{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y)}_{-H(X|Y)} - \sum_{x\in\mathcal{X}}\underbrace{\sum_{y\in\mathcal{Y}} p(x,y)}_{p(x)}\log p(x) \\ &= -H(X|Y) - \underbrace{\sum_{x\in\mathcal{X}} p(x)\log p(x)}_{H(X)} = H(X) - H(X|Y) \end{aligned}$$

Symmetrically, we also have $I(X;Y) = H(Y) - H(Y|X)$. Since $H(X,Y) = H(X) + H(Y|X)$ by chain rule, we have $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(X,Y) + H(X)$. $I(X;Y) = I(Y;X)$ by the symmetry of the formula.

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

□

From the equation $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, we can also intuitively see that mutual information represents the reduction of uncertainty of one random variable due to the knowledge of the other one's.

Example 1: Binary Noise Channel Consider the binary noise channel example (Example 2 in Section 1.3) again.

The mutual information of X and Y is

$$I(X;Y) = H(X) - H(X|Y) = 1 - h(\delta)$$

when $\delta = 0, 1$, $I(X;Y) = 1$ bits, thus after knowing Y, we have known all the information about X. when $\delta = 1/2$, $I(X;Y) = 0$, the channel Y is completely noisy with no information provided.

Below we define the conditional mutual information.

Definition 1.5.3: Conditional Mutual Information

The **conditional mutual information** of r.v. X and Y given Z is the reduction in the uncertainty of X due to knowledge of Y when Z is given.

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}_{p(x,y,z)}\left[\log\left(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right)\right]$$

The chain rule for conditional mutual information is below.

Theorem 1.5.4: Chain Rule for Mutual Information

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

Proof. We didn't define mutual information between multiple r.v.s and one r.v.. However, the definition is then trivial if thinking $X_1, \cdots, X_n$ as one random vector. Then the proof is as follow:

$$\begin{aligned} I(X_1, \cdots, X_n; Y) &= H(X_1, \cdots, X_n) - H(X_1, \cdots, X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1, Y) && \text{(Chain Rule for Entropy)} \\ &= \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1) \end{aligned}$$

which shows the original equation. □

Example 2: Let

$$(X, Y, Z) = \begin{cases} (0,0,0), & \text{with probability } 1/4 \\ (0,1,1), & \text{with probability } 1/4 \\ (1,0,1), & \text{with probability } 1/4 \\ (1,1,0), & \text{with probability } 1/4 \end{cases}$$

Then, we have

$$I(X;Y) = H(X) - H(X|Y) = 1 - 1 = 0 \text{ bits}$$

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 1 - 0 = 1 \text{ bits}$$

Chapter 2

Properties of Information Measures

2.1 Information Inequalities

Theorem 2.1.1: Information Inequality

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

$$D(p\|q) \geqslant 0 \quad \text{with equality iff } p(x) = q(x) \text{ for all } x$$

Proof. Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Then,

$$\begin{aligned} -D(p\|q) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} = \mathbb{E}_p\left[\log\frac{q(x)}{p(x)}\right] \\ &\leqslant \log \mathbb{E}_p\left[\frac{q(x)}{p(x)}\right] = \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} && \text{(Jensen's Inequality)} \\ &= \log\sum_{x\in A} q(x) \leqslant \log\sum_{x\in\mathcal{X}} q(x) = \log 1 = 0 \end{aligned}$$

The Jensen's inequality's equality condition is that, $X = c$, where c is a constant, with probability 1. In our case, $q(x)/p(x) = c$ for all x. Thus, $q(x) = cp(x)$. To satisfy $\sum_{x\in\mathcal{X}} q(x) = 1$, we must have $c = 1$ since $\sum_{x\in\mathcal{X}} p(x) = 1$. This implies that $p(x) = q(x)$ for all x. □

Corollary 2.1.2: Nonnegativity of other Measures

1. $I(X;Y) \geqslant 0$ with equality iff X, Y independent.
2. $D(p(y|x)\|q(y|x)) \geqslant 0$ with equality iff $p(y|x) = q(y|x)$ for all y and x such that $p(x) > 0$.
3. $I(X;Y|Z) \geqslant 0$ with equality iff X, Y conditionally independent given Z.

Any random variable has no greater entropy than the uniform one.

Proposition 2.1.3: Maximum Entropy

$H(X) \leqslant \log|\mathcal{X}|$, with equality iff X is uniformly distributed over $\mathcal{X}$.

Proof. Let $u(x) = 1/|\mathcal{X}|$ be the uniform pmf over $\mathcal{X}$. Let $p(x)$ be the pmf of X. Then,

$$D(p\|u) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} - \left(-\sum_{x\in\mathcal{X}} p(x)\log p(x)\right) = \log|\mathcal{X}| - H(X) \geqslant 0$$

which shows the result by using information inequality. □

Conditioning reduces entropy.

Proposition 2.1.4: Information can't hurt

$H(X|Y) \leqslant H(X)$ with equality iff X,Y independent.

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

Remark: This can only be true on average. Indeed, $H(X|Y=y)$ may be greater than $H(X)$, but on average $H(X|Y)$ must be less than $H(X)$. For example, suppose

$$(X,Y) = \begin{cases} (1,2), & \text{with probability } 1/8 \\ (2,1), & \text{with probability } 3/4 \\ (2,2), & \text{with probability } 1/8 \end{cases}$$

Then, $H(X) = 0.544$ bits, $H(X|Y=0) = 0$ bits, and $H(X|Y=2) = 1$ bits. Thus, $H(X|Y=2) \geqslant H(X)$. However, the average $H(X|Y) = 0.25$ bits $\leqslant H(X)$.

Corollary 2.1.5: Information can't hurt (Generalization)

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

with equality iff X_i's are independent.

Proof. By chain rule for entropy,

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

equality iff X_i is independent of $X_{i-1},\cdots,X_1$, i.e., X_i's are independent. □

Proposition 2.1.6: Entropy of Functions of a Random Variable

Let X be discrete r.v. Then, $H(g(X)) \leqslant H(X)$ for arbitrary function g.

Proof. On one hand,

$$\begin{aligned} H(X, g(X)) &= H(X) + H(g(X)|X) && \text{(Chain rule)} \\ &= H(X) + \sum_{x \in \mathcal{X}} p(x) H(g(X)|X = x) = H(X) && (g(X)|X = x \text{ is deterministic}) \end{aligned}$$

On the other hand,

$$\begin{aligned} H(X, g(X)) &= H(g(X)) + H(X|g(X)) && \text{(Chain rule)} \\ &\geqslant H(g(X)) && \text{(Nonnegativity of entropy)} \end{aligned}$$

Combining these two, we have $H(g(X)) \leqslant H(X)$. □

2.2 Convexity and Concavity of Information Measures

We start with a lemma.

Lemma 2.2.1: Log Sum Inequality

For nonnegative numbers $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$, we have

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geqslant \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i} = \text{const}$ for all i. We use the convention that $0 \log 0 = 0, a \log \frac{a}{0} = \infty$, $0 \log \frac{0}{b} = 0$ and $0 \log \frac{0}{0} = 0$.

Proof. Assume WLOG, $a_i > 0$ and $b_i > 0$. Since $f(t) = t \log t$ is strictly convex, we have

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geqslant f\left(\sum_{i=1}^{n} \alpha_i t_i \right)$$

where $\alpha_i \geqslant 0$ and $\sum_{i=1}^{n} \alpha_i = 1$. If we set $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $t_i = \frac{a_i}{b_i}$, we obtain

$$\text{LHS} = \sum_{i=1}^{n} \alpha_i f(t_i) = \sum_{i=1}^{n} \frac{b_i}{\sum_{j=1}^{n} b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} b_j} \log \frac{a_i}{b_i} = \frac{1}{\sum_{j=1}^{n} b_j} \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}$$

and

$$\text{RHS} = \left(\sum_{i=1}^{n} \frac{b_i}{\sum_{j=1}^{n} b_j} \frac{a_i}{b_i} \right) \log \left(\sum_{i=1}^{n} \frac{b_i}{\sum_{j=1}^{n} b_j} \frac{a_i}{b_i} \right) = \frac{1}{\sum_{j=1}^{n} b_j} \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{j=1}^{n} b_j}$$

Multiply both sides by $\sum_{j=1}^{n} b_j$, we get the log sum inequality. □

Relative entropy is convex in (p, q).

Theorem 2.2.2: Convexity of Relative Entropy

$D(p||q)$ is convex in (p, q). i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of pmfs, then

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leqslant \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2), \quad \forall\, 0 \leqslant \lambda \leqslant 1$$

Proof. Apply log sum inequality,

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

where we set $a_1 = \lambda p_1(x)$, $a_2 = (1-\lambda)p_2(x)$, $b_1 = \lambda q_1(x)$ and $b_2 = (1-\lambda)q_2(x)$. Sum over x, the result follows. □

Entropy is concave in p.

Theorem 2.2.3: Concavity of Entropy

$H(p)$ is a concave function of p.

Proof. Take u the uniform distribution on 2 outcomes, i.e., $\mathcal{X} = \{0, 1\}$, then

$$D(p||u) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{1/2} = \sum_{x\in\mathcal{X}} p(x) \log p(x) + \sum_{x\in X} p(x) \log_2 2 = -h(p) + 1$$

Since $D(p||u)$ is convex, $h(p)$ is concave. □

Theorem 2.2.4: Concave-Convexity of Mutual Information

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is

- Concave in $p(x)$ for fixed $p(y|x)$, and
- Convex in $p(y|x)$ for fixed $p(x)$.

Proof. To prove the first part, expand the mutual information

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

If $p(y|x)$ is fixed, then $p(y) = \sum_x p(y|x)p(x)$ is a linear function of $p(x)$ by law of total probability. Hence, $H(Y)$, which is a concave function of $p(y)$ by Theorem 2.2.3, is a concave function of $p(x)$. The second term is a linear function of $p(x)$. Hence, the difference is a concave function of $p(x)$.

To prove the second part, fix $p(x)$ and consider two different conditional distributions $p_1(y|x)$ and $p_2(y|x)$. The corresponding joint distributions are $p_1(x, y) = p(x)p_1(y|x)$ and $p_2(x, y) = p(x)p_2(y|x)$, and their respective marginals are $p(x), p_1(y)$ and $p(x), p_2(y)$. Consider the conditional distribution

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x), \quad \lambda \in [0, 1]$$

In this case, the distribution of Y is also a mixture

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda)p_2(y)$$

Hence, if we let $q_\lambda(x,y) = p(x)p_\lambda(y)$ be the product of the marginals, we have

$$q_\lambda(x,y) = \lambda q_1(x,y) + (1-\lambda)q_2(x,y)$$

Then the mutual information can be written as

$$I(X;Y) = D(p_\lambda(x,y)\|q_\lambda(x,y))$$

since relative entropy is convex in (p,q), it follows that the mutual information is convex w.r.t. the conditional distribution. □

2.3 Data-Processing Inequality and Sufficient Statistics

The data-processing inequality can be used to show that no clever manipulation of the data can improve the inference that can be made from the data.

Definition 2.3.1: Markov Chain

R.v.s X, Y, Z are said to form a **Markov Chain** $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is conditionally independent of X, i.e.,

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

Remarks:

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y. i.e., $p(x,z|y) = p(x|y)p(z|y)$.
- $X \to Y \to Z$ implies $Z \to Y \to X$.
- If $Z = f(Y)$, then $X \to Y \to f(Y)$.

Theorem 2.3.2: Data-Processing Inequality

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

Proof. By chain rule,

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

However, $I(X;Z|Y) = 0$ since X and Z are conditionally independent given Y. Thus, we have

$$I(X;Y) = I(X;Z) + I(X;Y|Z) \geqslant I(X;Z)$$

with equality iff $I(X;Y|Z) = 0$, i.e., $X \to Z \to Y$. Similarly, $I(Y;Z) \geqslant I(X;Z)$. □

Corollary 2.3.3: Data-Processing Inequality (Other Forms)

- If $Z = g(Y)$, we have $I(X;Y) \geqslant I(X;g(Y))$.
- If $X \to Y \to Z$, then $I(X;Y|Z) \leqslant I(X;Y)$.

Proof.

- The first directly follows from the fact that $X \to Y \to g(Y)$ forms a Markov Chain.
- Since $I(X;Y) = I(X;Z) + I(X;Y|Z) \geqslant I(X;Y|Z)$, by proof of Data-processing inequality.

□

Remark: If X, Y, Z does not form a Markov chain, the inequality can be not true. For example, let $X, Y \sim \text{Ber}(1/2)$ independent, and $Z = X + Y$. Then, $I(X;Y) = 0$, but $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = \mathbb{P}(Z = 1)H(X|Z=1) + P(Z=0)H(X|Z=0) = 1/2 + 0 = 1/2$ bits.

Suppose we have a family of pmfs $\{f_\theta\}$ indexed by θ, and let X be a sample from a distribution in this family. Let $T(X)$ be any statistic. Then, $\theta \to X \to T(X)$. By data-processing inequality, $I(\theta;T(X)) \leqslant I(\theta;X)$ for any distribution of θ. However, if equality holds, no information is lost. This leads to the *sufficient statistics*, where it contains all information in X about θ.

Definition 2.3.4: Sufficient Statistics

A function $T(X)$ is said to be a **sufficient statistics** relative to the family $\{f_\theta\}$ if X is independent of θ given $T(X)$ for any distribution on θ, i.e., $\theta \to T(X) \to X$ forms a Markov Chain. i.e.,

$$I(\theta;X) = I(\theta;T(X))$$

Definition 2.3.5: Minimal Sufficiency

A statistic $T(X)$ is a **minimal sufficient statistic** realtive to $\{f_\theta\}$ if it is a function of every other sufficient statistic U. i.e.,

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

2.4 Fano's Inequality

Suppose we observe a r.v. Y, and we wish to guess the value of a correlated r.v. X. To do this, we calculate a function $g(Y) = \hat{X}$. We wish to bound the probability of error, i.e., probability of $\hat{X} \neq X$. $X \to Y \to g(Y) = \hat{X}$ is a Markov Chain, and let $P_e = \mathbb{P}[\hat{X} \neq X]$.

Theorem 2.4.1: Fano's Inequality

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, we have

$$h(P_e) + P_e \log|\mathcal{X}| \geqslant H(X|\hat{X}) \geqslant H(X|Y)$$

This inequality can be weakened to

$$1 + P_e \log|\mathcal{X}| \geqslant H(X|Y) \quad \Longrightarrow \quad P_e \geqslant \frac{H(X|Y) - 1}{\log|\mathcal{X}|}$$

Proof. Define the error random variable

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

Then, using the chain rule for entropy,

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = H(E|\hat{X}) + H(X|E, \hat{X}) \tag{2.1}$$

First, since conditioning reduces entropy, we have

$$H(E|\hat{X}) \leqslant H(E) = h(P_e) \tag{2.2}$$

Second, we can bound the other term like

$$H(X|E, \hat{X}) = \mathbb{P}[E = 0]H(X|\hat{X}, E = 0) + \mathbb{P}[E = 1]H(X|\hat{X}, E = 1) \leqslant (1 - P_e) \times 0 + P_e \times \log|\mathcal{X}| \tag{2.3}$$

since given $E = 0$, we have $X = \hat{X}$, and given $E = 1$, we can upper bound the conditional entropy by the log of the number of possible outcomes by Proposition 2.1.3. Combining Equation 2.1, 2.2 and 2.3, we have

$$H(X|\hat{X}) \leqslant h(P_e) + P_e \log|\mathcal{X}|$$

By data-processing inequality, we have $I(X; \hat{X}) \leqslant I(X; Y)$ since $X \to Y \to \hat{X}$ is a Markov Chain. This shows that

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \leqslant I(X; Y) = H(X) - H(X|Y) \quad \Longrightarrow \quad H(X|\hat{X}) \geqslant H(X|Y)$$

and this completes the proof. The weakened inequality just follows from $h(\cdot) \leqslant 1$ bit. □

We can strengthen the inequality in two ways.

Corollary 2.4.2: Modification of Fano's Inequality

- For any two r.v.s X and Y, let $p = \mathbb{P}[X \neq Y]$, we have

$$h(p) + p \log|\mathcal{X}| \geqslant H(X|Y)$$

- Besides the conditions of Fano's inequality, if we assume that $g(Y) = \hat{X}$ has the same support with X, i.e., $\hat{X} : \mathcal{Y} \to \mathcal{X}$, we have

$$h(P_e) + P_e \log(|\mathcal{X}| - 1) \geqslant H(X|Y)$$

Proof. For the first one, just let $\hat{X} = Y$ in Fano's inequality. For the second one, we change the bound in Equation 2.3

by

$$H(X|E,\hat{X}) = \mathbb{P}[E=0]H(X|\hat{X},E=0) + \mathbb{P}[E=1]H(X|\hat{X},E=1) \leqslant (1-P_e)\times 0 + P_e \times (\log|\mathcal{X}|-1)$$

since it excludes one point $\hat{X} = X$ out of the support. □

Remark: The Fano's Inequality in Corollary is sharp

Suppose that there is no knowledge of Y. Thus, X must be guessed without any information. Let $X \in \{1,2,\cdots,m\}$ and $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m$. Then, the best guess is $\hat{X} = 1$, the resulting probability of error is $P_e = 1 - p_1$. Fano's inequality then becomes

$$h(P_e) + P_e \log(m-1) \geqslant H(X)$$

To achieve the equality, we assume the pmf

$$(p_1,p_2,\cdots,p_m) = \left(1-P_e, \frac{P_e}{m-1}, \cdots, \frac{P_e}{m-1}\right)$$

Then, we have

$$h(P_e) + P_e\log(m-1) = -P_e\log P_e - (1-P_e)\log(1-P_e) + P_e\log(m-1) = -P_e\log\frac{P_e}{m-1} - (1-P_e)\log(1-P_e) = H(X)$$

2.5 Inequalities Relating Probability of Error and Entropy

Lemma 2.5.1: Probability of Error Bound of i.i.d. r.v.s

Let X and X' be i.i.d. with entropy $H(X)$, then

$$\mathbb{P}[X = X'] \geqslant 2^{-H(X)}$$

with equality iff X has a uniform distribution.

Proof. Suppose that $X \sim p(x)$. By Jensen's inequality, we have

$$2^{\mathbb{E}[\log p(X)]} \leqslant \mathbb{E}\left[2^{\log p(X)}\right]$$

which implies that

$$2^{-H(X)} = 2^{\sum p(x)\log p(x)} \leqslant \sum p(x) 2^{\log p(x)} = \sum p^2(x) = \mathbb{P}[X = X']$$

□

Lemma 2.5.2: Probability of Error Bound of independent Same Support r.v.s

Let $X \sim p(x)$ and $X' \sim r(x)$, $x, x' \in \mathcal{X}$ be independent. Then,

$$\mathbb{P}[X = X'] \geqslant 2^{-H(p)-D(p||r)}$$

$$\mathbb{P}[X = X'] \geqslant 2^{-H(r)-D(r||p)}$$

Proof.

$$\begin{aligned}
2^{-H(p)-D(p||r)} &= 2^{\sum p(x)\log p(x)+\sum p(x)\log\frac{r(x)}{p(x)}} \\
&= 2^{\sum p(x)\log r(x)} \\
&\leqslant \sum p(x)2^{\log r(x)} && \text{(Jensen's Inequality)} \\
&= \sum p(x)r(x) = \mathbb{P}[X = X']
\end{aligned}$$

□

Chapter 3

Asymptotic Equipartition Property

'Almost all events are almost equally surprising'.

3.1 Asymptotic Equipartition Property Theorem

Theorem 3.1.1: AEP

If $X_1, X_2, \cdots$ are i.i.d. r.v.s with density $p(x)$, then

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

Proof. Functions of independent r.v.s are also independent. Thus, $\log p(X_i)$ are independent. By week law of large numbers,

$$\begin{aligned} -\frac{1}{n}\log p(X_1, \cdots, X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \\ &- \mathbb{E}[\log p(X)] \quad \text{in probability} && \text{(WLLN)} \\ &= H(X) \end{aligned}$$

which proves the theorem. □

Definition 3.1.2: Typical Set

The **typical set** $A_\epsilon^{(n)}$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ with the property

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

Below are properties of typical set.

Proposition 3.1.3: Properties of Typical Sets

1. *All elements nearly equiprobable*: If $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leqslant -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) \leqslant H(X) + \epsilon$.
2. *Probability nearly 1*: $\mathbb{P}\left[A_\epsilon^{(n)}\right] > 1 - \epsilon$ for n sufficiently large.
3. *Number of elements nearly 2^{nH} I*: $|A_\epsilon^{(n)}| \leqslant 2^{n(H(X)+\epsilon)}$
4. *Number of elements nearly 2^{nH} II*: $|A_\epsilon^{(n)}| \geqslant (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Proof. 1. Immediate from definition.

2. By Theorem 3.1.1, the probability of $(X_1, X_2, \cdots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. Thus, for any $\delta > 0$, there exists an $n(\delta, \epsilon)$ such that for all $n \geqslant n(\delta, \epsilon)$, we have

$$\mathbb{P}\left[\left|-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) - H(X)\right| < \epsilon\right] > 1 - \delta$$

Setting $\delta = \epsilon$, we obtain the second part of the theorem.

3. We have

$$\begin{aligned} 1 = \sum_{(x_1,\cdots,x_n)\in\mathcal{X}^n} p(x_1, \cdots, x_n) &\geqslant \sum_{(x_1,\cdots,x_n)\in A_\epsilon^{(n)}} p(x_1, \cdots, x_n) \\ \geqslant \sum_{(x_1,\cdots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} &= 2^{-n(H(X)+\epsilon)}\left|A_\epsilon^{(n)}\right| \end{aligned}$$

where the inequality of the second row follows from the definition.

4. For sufficiently large n, $\mathbb{P}\left[A_\epsilon^{(n)}\right] > 1 - \epsilon$, so

$$1 - \epsilon < \mathbb{P}\left[A_\epsilon^{(n)}\right] \leqslant \sum_{(x_1,\cdots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)}\left|A_\epsilon^{(n)}\right|$$

which completes the proof.

□

3.2 AEP and Data Compression

Let $X_1, X_2, \cdots, X_n$ be i.i.d. r.v.s drawn from pmf $p(x)$. We wish to find short coding for such sequence of r.v.s. We divide all sequences in $\mathcal{X}^n$ into two sets: the typical set $A_\epsilon^{(n)}$ and its complement. We order all elements in each set according to some order. Then, we can represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set.

- Since there are less than or equal to $2^{n(H+\epsilon)}$ sequences in the typical set, the indexing requires no more than $n(H + \epsilon) + 1$ bits (+1 because $n(H + \epsilon)$ may not be an integer, take $\log_2$ because we code using binary digits). We prefix all these sequences by a 0, giving a total length of $\leqslant n(H + \epsilon) + 2$ bits to represent each sequence in $A_\epsilon^{(n)}$.

- Similarly, we index each sequence not in $A_\epsilon^{(n)}$ by using not more than $n \log |\mathcal{X}|+1$ bits by brute-force enumeration. Prefix these indices by 1, giving a total length of $\leqslant n \log |\mathcal{X}|+2$ bits.

Using this machinery, we can represent sequence $(X_1, \cdots, X_n)$ using $nH(X)$ bits on the average.

Theorem 3.2.1: Average Coding Length

Let $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$. Let $\epsilon > 0$. Then, there exists a code that maps sequence $(x_1, \cdots, x_n)$ of length n into binary strings of length $\ell(x_1, \cdots, x_n)$ such that the mapping is one-to-one (invertible) and

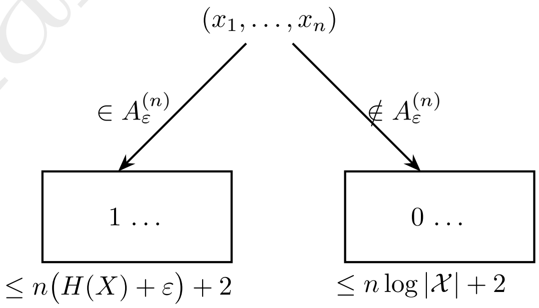
$$\mathbb{E}\left[\frac{1}{n}\ell(X_1, \cdots, X_n)\right] \leqslant H(X) + \epsilon$$

for n sufficiently large.

Proof. If n is sufficiently large, we have $\mathbb{P}\left[A_\epsilon^{(n)}\right] > 1-\epsilon$. Then, the expected length of codeword is

$$\begin{aligned}
\mathbb{E}[\ell(X_1, \cdots, X_n)] &= \sum_{x_1, \cdots, x_n} p(x_1, \cdots, x_n)\ell(x_1, \cdots, x_n) \\
&= \sum_{(x_1, \cdots, x_n) \in A_\epsilon^{(n)}} p(x_1, \cdots, x_n)\ell(x_1, \cdots, x_n) + \sum_{(x_1, \cdots, x_n) \notin A_\epsilon^{(n)}} p(x_1, \cdots, x_n)\ell(x_1, \cdots, x_n) \\
&\leqslant \sum_{(x_1, \cdots, x_n) \in A_\epsilon^{(n)}} p(x_1, \cdots, x_n)(n(H+\epsilon)+2) + \sum_{(x_1, \cdots, x_n) \notin A_\epsilon^{(n)}} p(x_1, \cdots, x_n)(n \log |\mathcal{X}|+2) \\
&= \mathbb{P}\left[A_\epsilon^{(n)}\right](n(H+\epsilon)+2) + \left(1-\mathbb{P}\left[A_\epsilon^{(n)}\right]\right)(n \log |\mathcal{X}|+2) \\
&\leqslant n(H+\epsilon)+2+\left(1-\mathbb{P}\left[A_\epsilon^{(n)}\right]\right)(n \log |\mathcal{X}|+2) && \left(\mathbb{P}\left[A_\epsilon^{(n)}\right] \leqslant 1\right) \\
&\leqslant n(H+\epsilon)+2+\epsilon n \log |\mathcal{X}| && \left(\mathbb{P}\left[A_\epsilon^{(n)}\right] > 1-\epsilon\right) \\
&= n(H+\tilde{\epsilon})
\end{aligned}$$

where $\tilde{\epsilon} = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$. This can be made arbitrarily small with sufficiently small ϵ and large n. □


$(x_1, \ldots, x_n)$
$\in A_\varepsilon^{(n)}$
$\notin A_\varepsilon^{(n)}$
$1 \ldots$
$0 \ldots$
$\leq n\big(H(X)+\varepsilon\big)+2$
$\leq n \log |\mathcal{X}|+2$

Chapter 4

Entropy Rates of a Stochastic Process

4.1 Markov Chains

Definition 4.1.1: Stationary Process

A stochastic process is said to be **stationary** if the joint distribution of any subset of the sequence of r.v.s is invariant w.r.t. shifts in the time index, i.e.,

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] = \mathbb{P}[X_{1+\ell} = x_1, X_{2+\ell} = x_2, \cdots, X_{n+\ell} = x_n]$$

for every n and every shift l, and for all $x_1, x_2, \cdots, x_n \in \mathcal{X}$.

Definition 4.1.2: Markov Chain

A discrete stochastic process $X_1, X_2, \cdots$ is said to be a **Markov Chain** (or **Markov Process**) if for $n = 1, 2, \cdots$,

$$\mathbb{P}[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_1 = x_1] = \mathbb{P}[X_{n+1} = x_{n+1} | X_n = x_n]$$

for all $x_1, x_2, \cdots, x_n, x_{n+1} \in \mathcal{X}$. In this case, the joint pmf can be written as

$$p(x_1, \cdots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_n|x_{n-1})$$

Definition 4.1.3: Time Invariant Markov Chain

A Markov chain is said to be **time invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n, that is, for $n = 1, 2, \cdots$,

$$\mathbb{P}[X_{n+1} = b | X_n = a] = \mathbb{P}[X_2 = b | X_1 = a], \quad \forall\, a, b \in \mathcal{X}$$

We assume Markov chain is time invariant unless otherwise stated.

A time-invariant Markov chain is characterized by

- Initial state X_1,
- Probability transition matrix $P_{ij} = \mathbb{P}[X_{n+1} = j | X_n = i]$

Definition 4.1.4: Irreducible/Aperiodic

- A Markov chain is **irreducible** if it is poosible to go with positive probability from any state to any other state in a finite number of steps.
- A Markov chain is **aperiodic** if the gcd of the length of different path from a state to itself is 1.

If the pmf of the r.v. at time n is $p(x_n)$, the probability mass function at time $n+1$ is

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n, x_{n+1}}$$

A distribution on the state such that the distribution at time $n+1$ is the same as the distribution at time n is called a **stationary distribution**. The stationary distriubtion μ saitsifies $\mu P = \mu$. *If the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain forms a stationary process.*

If the finite-state Markov chain is irreducible and aperiodic, the *stationary distribution* is **unique**, and from any starting distribution, the distribution of X_n tends to the stationary distribution as $n \to \infty$.

Example: Consider a two-state Markov chain with a probability transition matrix

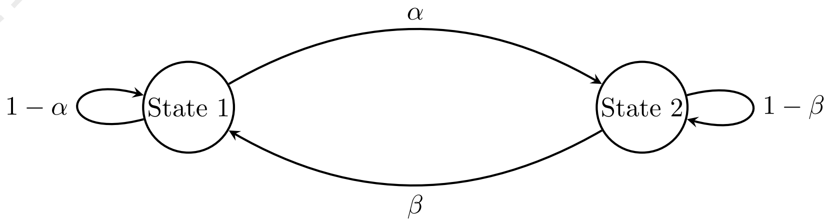
$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

Thus,

$$\begin{pmatrix} \mu_1 & \mu_2 \end{pmatrix} \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} = \begin{pmatrix} (1-\alpha)\mu_1 + \beta\mu_2 & \alpha\mu_1 + (1-\beta)\mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 & \mu_2 \end{pmatrix} \quad \Longrightarrow \quad \mu_1 \alpha = \mu_2 \beta$$

Since $\mu_1 + \mu_2 = 1$, we have

$$\mu_1 = \frac{\beta}{\alpha+\beta}, \quad \mu_2 = \frac{\alpha}{\alpha+\beta}$$



The entropy of the state X_n at time n is $H(X_n) = h\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$, and this is not the rate at which entropy grows for $H(X_1, X_2, \cdots, X_n)$.

4.2 Entropy Rate

Definition 4.2.1: Entropy of Stochastic Process

The **entropy** of a stochastic process $\mathbb{X} = \{X_i\}$ is defined by

$$H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

provided the limit exists.

Example 1: Typewriter Consider the typewriter that has m equally likely output letters. The typewriter can produce m^n sequences of length n, all of them are equally likely. Hence, $H(X_1, \cdots, X_n) = \log m^n$, and the entropy rate is $H(\mathbb{X}) = \log m$ bits per symbol.
Example 2: i.i.d. Samples Let $X_1, X_2, \cdots$ be i.i.d. r.v.s. Then,

$$H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n) = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1)$$

Example 3: Independent Samples Let $X_1, X_2, \cdots$ be independent but not indentically distributed. Then,

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

$H(X_i)$'s are not equal. We can choose a sequence of distribution such that the limit of $\frac{1}{n}\sum_{i=1}^{n} H(X_i)$ does not exist. For example, let X_i's are all Bernoulli r.v.s with success probability p_i dependent on i, then choose p_i such that

$$p_i = \begin{cases} 0.5, & \text{if } 2k < \log(\log i) \leqslant 2k+1, \\ 0, & \text{if } 2k+1 < \log(\log i) \leqslant 2k+2 \end{cases}, k = 0, 1, 2, \cdots$$

Then, $H(X_i) = 1$ when $p_i = 0.5$ and $H(X_i) = 0$ when $p_i = 0$. Then, it will oscillate between nearly 0 and nearly 1 with larger and larger blocks. Thus, it does not converge.

We first show that per symbol entropy is equal to the conditional entropy of the last r.v. for stationary processes.

Theorem 4.2.2: Equivalence Between Two Entropy Rate

For a stationary stochastic process, the limits

$$H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

and

$$H'(\mathbb{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \cdots, X_1)$$

both exist, and they are equal:

$$H(\mathbb{X}) = H'(\mathbb{X})$$

Proof. We first prove that $\lim H(X_n|X_{n-1},\cdots,X_1)$ exists. To do this, we note that

$$\begin{aligned} H(X_{n+1}|X_1,\cdots,X_n) &\leqslant H(X_{n+1}|X_n,\cdots,X_2) && \text{(Conditioning decreases entropy)} \\ &= H(X_n|X_{n-1},\cdots,X_1) && \text{(Stationary Process)} \end{aligned}$$

Therefore, $H(X_n|X_{n-1},\cdots,X_1)$ is a decreasing sequence of nonnegative numbers, the limit exists.

Now by the chain rule,

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

By *Cesáro mean* (If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$), LHS also tends to the same limit $H'(\mathbb{X})$ as RHS. □

4.3 Entropy Rate of Markov Chain

For a stationary Markov chain, the entropy rate is given by

$$H(\mathbb{X}) = H'(\mathbb{X}) = \lim_{n\to\infty} H(X_n|X_{n-1},\cdots,X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1)$$

If $X_1 \sim \mu$, the stationary distribution, then the entropy rate is

$$\begin{aligned} H(\mathbb{X}) = H(X_2|X_1) &= -\sum_{x_1,x_2} p(x_1)p(x_2|x_1)\log p(x_2|x_1) \\ &= -\sum_i \mu_i \sum_j P_{ij}\log P_{ij} = -\sum_{ij} \mu_i P_{ij}\log P_{ij} \end{aligned}$$

Example1 : We continue the example from section 4.1, the entropy rate is

$$H(\mathbb{X}) = H(X_2|X_1) = -\sum_{i,j=1,2} \mu_i P_{ij}\log P_{ij} = \frac{\beta}{\alpha+\beta}h(\alpha) + \frac{\alpha}{\alpha+\beta}h(\beta)$$

Remark: If the Markov chain is irreducible and aperiodic, the stationary distribution is unique, and distribution tends to stationary as $n \to \infty$. In this case, even though the initial distribution is not the stationary distribution, the entropy rate is still $H(\mathbb{X})$ since it only concerns the long term behavior.

Example 2: Entropy Rate of a Random Walk on a Stochastic Graph Consider a graph of m nodes $\{1,2,\cdots,m\}$ with weighted edges ω_{ij}, and undirected ($\omega_{ij} = \omega_{ji}$, and $\omega_{ij} = 0$ iff no edge between i and j). Consider the random walk $\{X_n\}$, $X_n \in \{1,2,\cdots,m\}$ and transition matrix

$$P_{ij} = \frac{\omega_{ij}}{\omega_i}, \quad \omega_i = \sum_{j=1}^{m} \omega_{ij}$$

In this case, the stationary distribution is

$$\mu = (\mu_1,\cdots,\mu_m), \quad \mu_i = \frac{\omega_i}{2\omega}, \quad \omega = \sum_{i,j:i>j} \omega_{ij}$$

To verify that it is indeed a stationary distribution,

$$\sum_{i=1}^{m} \mu_i P_{ij} = \sum_{i=1}^{m} \frac{\omega_i}{2\omega}\frac{\omega_{ij}}{\omega_i} = \frac{\omega_j}{2\omega} = \mu_j \quad\Longrightarrow\quad \mu P = \mu$$

The entropy rate is

$$\begin{aligned}
H(\mathbb{X}) = H(X_2|X_1) &= -\sum_{i=1}^{m}\mu_i \sum_{j=1}^{m} P_{ij}\log P_{ij} \\
&= -\sum_{i=1}^{m}\frac{\omega_i}{2\omega}\sum_{j=1}^{m}\frac{\omega_{ij}}{\omega_i}\log\left(\frac{\omega_{ij}}{\omega_i}\right) = -\sum_{i,j}\frac{\omega_i}{2\omega}\frac{\omega_{ij}}{\omega_i}\log\left(\frac{\omega_{ij}}{\omega_i}\cdot\frac{2\omega}{2\omega}\right) \\
&= -\sum_{i,j}\frac{\omega_{ij}}{2\omega}\log\left(\frac{\omega_{ij}}{2\omega}\right) - \left(-\sum_{i,j}\frac{\omega_{ij}}{2\omega}\log\left(\frac{\omega_i}{2\omega}\right)\right) \\
&= -\sum_{i,j}\frac{\omega_{ij}}{2\omega}\log\left(\frac{\omega_{ij}}{2\omega}\right) - \underbrace{\left(-\sum_{i=1}^{m}\frac{\omega_i}{2\omega}\log\left(\frac{\omega_i}{2\omega}\right)\right)}_{H(\mu)}
\end{aligned}$$

If all weights are equal, the stationary distribution is then $\mu_i = E_i/2E$, where E_i is the number of edges connecting node i, E is the total number of edges. Then, the entropy rate is

$$H(\mathbb{X}) = \log(2E) - h\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \cdots, \frac{E_m}{2E}\right)$$

4.4 Information Theoretical Properties of Markov Chain

Proposition 4.4.1: Decreasing Relative Entropy

Let μ_n, μ_n' be two probability distributions on the state space of Markov chain at time n. Then, $D(\mu_n||\mu_n')$ is non-increasing with n.

Proof. Let $p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$ and $q(x_n, x_{n+1}) = q(x_n)r(x_{n+1}|x_n)$. Then,

$$\begin{aligned}
D(p(x_n, x_{n+1})\|q(x_n, x_{n+1})) &= D(p(x_n)\|q(x_n)) + \underbrace{D(r(x_{n+1}|x_n)\|r(x_{n+1}|x_n))}_{=0} && \text{(Chain rule)} \\
&= D(p(x_{n+1})\|q(x_{n+1})) + D(p(x_n|x_{n+1})\|q(x_n|x_{n+1})) && \text{(Chain rule)} \\
&\geqslant D(p(x_{n+1})\|q(x_{n+1}))
\end{aligned}$$

Thus, $D(p(x_n)\|q(x_n)) \geqslant D(p(x_{n+1})\|q(x_{n+1}))$. □

Corollary 4.4.2: Decreasing Relative Entropy with Stationary Distribution

If the Markov chain attains a stationary distribution μ, then $D(\mu_n\|\mu)$ is non-increasing with n.

Proof. Let $\mu_n' = \mu$ in the last proposition. □

Proposition 4.4.3: Increasing Entropy

Entropy non-decreases if stationary distribution is uniform.

Proof.

$$D(\mu_n\|\mu) = \sum_{x_n} \mu_n(x_n) \log \frac{\mu_n(x_n)}{1/|\mathcal{X}|} = \log |\mathcal{X}| - H(X_n)$$

Since $D(\mu_n\|\mu)$ is non-increasing, $H(X_n)$ is non-decreasing. □

Which Markov chain attains the uniform stationary distribution?

Definition 4.4.4: Doubly Stochastic

A transition matrix P is doubly stochastic if

$$\sum_i P_{ij} = 1, \quad \sum_j P_{ij} = 1$$

Lemma 4.4.5: Doubly stochastic and Uniform Stationary

The uniform distribution is a stationary distribution of P if and only if the transition matrix is doubly stochastic.

Proof. ($\Longrightarrow$) If m-dimensional P is doubly stochastic, then if $\mu_i = 1/m$, we have

$$\sum_{i=1}^{m} \mu_i P_{ij} = \frac{1}{m} \sum_i P_{ij} = \frac{1}{m} = \mu_j$$

Thus, uniform distribution is one of the stationary distribution.
($\Longleftarrow$) Suppose uniform is a stationary distribution, then

$$\frac{1}{m} = \mu_j = \sum_i \mu_i P_{ij} = \frac{1}{m} \sum_i P_{ij}$$

Thus, $\sum_i P_{ij} = 1$. Similarly, $\sum_j P_{ij} = 1$, the matrix is doubly stochastic. □

Proposition 4.4.6: Conditional Entropy increases

For a stationary Markov chain, $H(X_n|X_1)$ is non-decreasing with n.

Proof.

$$H(X_n|X_1) \geqslant H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_{n-1}|X_1)$$

where the inequality holds by conditioning reduces entropy, the first equality holds by Markov property, and the second equality holds by stationary.

Alternative proof:

$X_1 \to X_{n-1} \to X_n$, by data processing inequality,

$$H(X_{n-1}) - H(X_{n-1}|X_1) = I(X_1; X_{n-1}) \geqslant I(X_1; X_n) = H(X_n) - H(X_n|X_1)$$

the results follows form that $H(X_{n-1}) = H(X_n)$ by stationary. □

Proposition 4.4.7: Suffles Increases entropy

If T is a permutation of X, and the choice of the shuffle T is independent of X, then

$$H(TX) \geqslant H(X)$$

Proof.

$$H(TX) \geqslant H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

The inequality follows from the fact that conditioning reduces entropy, the first equality follows from that given T, we can reverse the shuffle, and the final equality follows that X and T are independent. □

4.5 Functions of Markov Chain

Let $X_1, X_2, \cdots$ be a stationary Markov chain. Let $Y_i = \phi(X_i)$. Y may not be a Markov chain, **but it is still stationary**.

Example: Let the states be $\{0, 1, 2\}$. Let the transition matrix be

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}$$

Let

$$Y_n = \begin{cases} 1, & X_n = 0 \\ 0, & X_n = 1, X_n = 2 \end{cases}$$

Then,

$$\mathbb{P}[Y_{n+1}|Y_n = 0, X_n = 1] = 0$$

$$\mathbb{P}[Y_{n+1} = 1|Y_n = 0, X_n = 2] = 1/2$$

it is not Markov chain.

However, we can get the upper and lower bounds.

Proposition 4.5.1: Upper Bound

$H(Y_n|Y_{n-1}, \cdots, Y_1)$ converges monotonically from above to $H(\mathbb{Y})$.

$$H(\mathbb{Y}) \leqslant H(Y_n|Y_{n-1}, \cdots, Y_1)$$

This is from previous section.

Proposition 4.5.2: Lower Bound

$$H(Y_n|Y_{n-1},\cdots,Y_2,X_1)\leqslant H(\mathbb{Y})$$

Proof.

$$\begin{aligned}
H(Y_n|Y_{n-1},\cdots,Y_2,X_1) &= H(Y_n|Y_{n-1},\cdots,Y_2,Y_1,X_1) && (Y_1 \text{ deterministic function of } X_1)\\
&= H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,\cdots,X_{-k}) && (\text{Markov property})\\
&= H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,\cdots,X_{-k},Y_0,\cdots,Y_{-k})\\
&\leqslant H(Y_n|Y_{n-1},\cdots,Y_1,Y_0,\cdots,Y_{-k}) && (\text{Conditioning reduces entropy})\\
&= H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1)
\end{aligned}$$

This holds for all $k=1,2,\cdots$, so

$$H(Y_n|Y_{n-1},\cdots,Y_2,X_1)\leqslant \lim_{k\to\infty} H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1)=H(\mathbb{Y})$$

where the equality holds since Y is stationary. □

Finally, the distance between upper and lower bound vanishes.

Proposition 4.5.3: Convergence of upper and lower bound

$H(Y_n|Y_{n-1},\cdots,Y_1)-H(Y_n|Y_{n-1},\cdots,Y_1,X_1)\to 0.$

Proof.

$$H(Y_n|Y_{n-1},\cdots,Y_1)-H(Y_n|Y_{n-1},\cdots,Y_1,X_1)=I(X_1;Y_n|Y_{n-1},\cdots,Y_1)$$

By the properties that

- $I(X_1;Y_1,\cdots,Y_n)=H(X_1)-H(X_1|Y_{n-1},\cdots,Y_1)\leqslant H(X_1)$
- Conditioning reduces entropy.

We have that $I(X_1;Y_n|Y_{n-1},\cdots,Y_1)$ increases with n, and is bounded by $H(X_1)$. Therefore,

$$H(X_1)\geqslant \lim_{n\to\infty} I(X_1;Y_1,\cdots,Y_n)=\lim_{n\to\infty}\sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\cdots,Y_1)=\sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\cdots,Y_1)$$

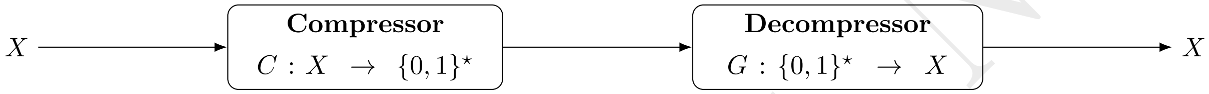
Since it is bounded above (i.e., converge), the tail must converge to 0, therefore,

$$\lim_{n\to\infty} I(X_1;Y_n|Y_{i-1},\cdots,Y_1)\to 0$$

□

Chapter 5

Data Compression


X
Compressor
$C : X \rightarrow \{0,1\}^\star$
Decompressor
$G : \{0,1\}^\star \rightarrow X$
X

5.1 Terminologies for Codes

Definition 5.1.1: Source Code/Codeword/Codeword Length

- A **source code** C for a r.v. X is a mapping from $\mathcal{X}$ to $\{0,1\}^\star$ (set of finite-length strings of symbols from a binary alphabet).
- $C(x) \in \{0,1\}^\star$ denotes the **codeword** of x, $\ell(x) = |C(x)|$ denotes the length of $C(x)$.

Example 1: $\mathcal{X} = \{\text{red}, \text{blue}\}$, $C(\text{red}) = 00$, $C(\text{blue}) = 11$.

Definition 5.1.2: Expected Length

The **expected length** $L(C)$ of a source code $C(x)$ for r.v. X with pmf $p(x)$ is given by

$$L(C) = \sum_{x\in\mathcal{X}} p(x)\ell(x)$$

Example 2: Let

$$\begin{cases} \mathbb{P}(X=1) = \frac{1}{2}, & C(1) = 0 \\ \mathbb{P}(X=2) = \frac{1}{4}, & C(2) = 10 \\ \mathbb{P}(X=3) = \frac{1}{8}, & C(3) = 110 \\ \mathbb{P}(X=4) = \frac{1}{8}, & C(4) = 111 \end{cases}$$

Then, we have expected length equals to entropy

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2 \times \frac{1}{8}\log 8 = 1.75 \text{ bits}$$

$$L(C) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 2 \times 3 = 1.75 \text{ bits} = H(X)$$

Example 3: Let

$$\begin{cases} \mathbb{P}(X=1) = \frac{1}{3}, & C(1) = 0 \\ \mathbb{P}(X=2) = \frac{1}{3}, & C(2) = 10 \\ \mathbb{P}(X=3) = \frac{1}{3}, & C(3) = 11 \end{cases}$$

Then, we have expected length larger than extropy:

$$H(X) = \log 3 = 1.58 \text{ bits}, \quad L(C) = \frac{1}{3} \times 1 + \frac{1}{3} \times 2 \times 2 = 1.66 \text{ bits} > H(X)$$

Definition 5.1.3: Non-singular Code

A code $C : \mathcal{X} \to \{0,1\}^\star$ is **non-singular** if $\forall\, x, x' \in \mathcal{X}$, $x \neq x' \implies C(x) \neq C(x')$.

However, non-singular code can be non-uniquely decoded (we can put ‘commas’ on difference places).

Definition 5.1.4: Extension

The **extension** $C^\star$ of a code $C : \mathcal{X} \to \{0,1\}^\star$ is

$$C^\star(x_1, \cdots, x_n) = C(x_1)C(x_2)\cdots C(x_n)$$

Example 4: For example 1, we have $C(\text{red}, \text{blue}) = 0011$.

Definition 5.1.5: Uniquely Decodable

A code $C : \mathcal{X} \to \{0,1\}^\star$ is **uniquely decodable** if its extension is non-singular.

However, we may still need future strings to determine some source strings. For example, if 3 is coded by 11, 4 is coded by 110, after we see string 11, we still need another one future string to see whether it is decoded as 3 or 4.

Definition 5.1.6: Prefix Code/Instantaneous Code

A code is called **prefix code** or **instantaneous code** if no codeword is a prefix of any other codeword.

An instantaneous code is “self-punctuating”.

X	Singular	Non-singular, but not uniquely decodable	Uniquely decodable, but not instantaneous	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

Table 5.1: Example of classes of codes

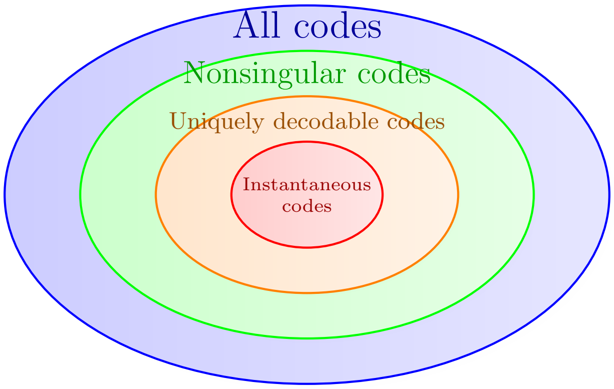


Figure 5.1: Relation among code structures, and examples

5.2 Kraft Inequality

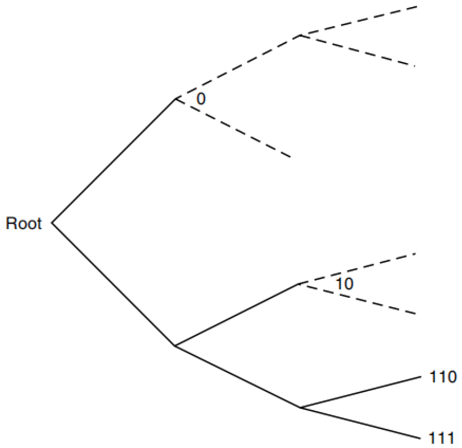
We wish to construct instantaneous codes of minimum expected length to describe a give source. The length of these kinds of code is limited by the following inequality.

Theorem 5.2.1: Kraft Inequality

Let $|\mathcal{X}| = m$. There exists a prefix code $C : \mathcal{X} \to \{0,1\}^\star$ with lengths $\ell_1, \ell_2, \cdots, \ell_m$ if and only if

$$\sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1$$

Proof. ($\Longrightarrow$) Suppose there exists a prefix code with length $\ell_1, \ell_2, \cdots, \ell_m$. Then, consider a binary tree of codewords.



Each of the codewords represents a path from root to a leaf node. The length of the path is ℓ_i, i.e., the length of the codeword. Since we need a prefix code, each codeword eliminates its descendants as possible codewords. These path are terminated as a result (dashed lines). Let $\ell_{\max}$ be the length of the longest codeword. A codeword at level ℓ_i has $2^{\ell_{\max}-\ell_i}$ ‘imaginary’ descendants at level $\ell_{\max}$. Consider all descendant of codewords. The sum of all these descendant cannot exceed the total number of seeds at level $\ell_{\max}$, i.e.,

$$\sum_{i=1}^{m} 2^{\ell_{\max}-\ell_i} \leqslant 2^{\ell\,\max} \quad \Longrightarrow \quad \sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1$$

($\Longleftarrow$) Conversely, given any set of codeword length $\ell_1, \ell_2, \cdots, \ell_m$ that satisfy the Kraft inequality, we can always construct a tree like the one in the Figure above. Label the first node of depth ℓ_1 as 1, and remove its descendants from the tree, then label the first remaining node of depth ℓ_2 as 2, and so on. □

We then show that uniquely decodable codes does not offer any further possibilities for the codeword length.

Theorem 5.2.2: McMillian: Kraft Inequality for Uniquely Decodable Codes

Let $|\mathcal{X}| = m$. There exists a uniquely decodable code $C : \mathcal{X} \to \{0,1\}^\star$ with length $\ell_1, \ell_2, \cdots, \ell_m$ if and only if

$$\sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1$$

Proof. ($\Longrightarrow$) Consider C^k, the kth extension of the code (i.e., the code formed by the concatenation of k repetitions of the given uniquely decodable code C). Let the codeword lengths of the symbols $x \in \mathcal{X}$ be denoted by $\ell(x)$. For the extension code, the length of the code sequence is

$$\ell(x_1, \cdots, x_k) = \sum_{i=1}^{k} \ell(x_i) \tag{5.1}$$

Now, we consider the quantity

$$\begin{aligned}\left(\sum_{x\in\mathcal{X}} 2^{-\ell(x)}\right)^k &= \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} 2^{-\ell(x_1)}2^{-\ell(x_2)}\cdots 2^{-\ell(x_k)} \\ &= \sum_{(x_1,x_2,\cdots,x_k)\in\mathcal{X}^k} 2^{-\ell(x_1)-\ell(x_2)-\cdots-\ell(x_k)} = \sum_{(x_1,x_2,\cdots,x_k)\in\mathcal{X}^k} 2^{-\ell(x_1,\cdots,x_k)}\end{aligned}$$

where the last step follows from the Equation 5.1. We now gather the terms by word length to obtain

$$\sum_{(x_1,\cdots,x_k)\in\mathcal{X}^k} 2^{-\ell(x_1,\cdots,x_n)} = \sum_{m=1}^{k\ell_{\max}} a(m)2^{-m}$$

where $\ell_{\max}$ is the maximum codeword length and $a(m)$ is the number of souce sequence $(x_1, \cdots, x_k)$ mapping into codewords of length m. Since the code is uniquely decodable, the kth extension of the code is nonsigular. Thus, there is at most one sequence mapping into each m-length code sequence. Together with the fact that there are at most 2^m binary code sequence of length m, we have $a(m) \leqslant 1 \times 2^m = 2^m$. Thus,

$$\left(\sum_{x\in\mathcal{X}} 2^{-\ell(x)}\right)^k = \sum_{m}^{k\ell_{\max}} a(m)2^{-m} \leqslant \sum_{m}^{k\ell_{\max}} 2^m 2^{-m} = k\ell_{\max}$$

Therefore, taking kth square root both sides, we have $\sum_{j=1}^{m} 2^{-\ell_j} \leqslant (k\ell_{\max})^{1/k}$. Since this inequality is true for all k, it is also true for $k \to \infty$. We have that

$$\lim_{k\to\infty} \frac{1}{k}\log(k\ell_{\max}) = \lim_{k\to\infty}\frac{1}{k} = 0 \quad\Longrightarrow\quad \lim_{k\to\infty}(k\ell_{\max})^{1/k} = 1$$

which leads to $\sum_{j=1}^{m} 2^{-\ell_j} \leqslant 1$, i.e., the Kraft inequality.

($\Longleftarrow$) Conversely, give any set of $\ell_1, \cdots, \ell_m$ satisfying the Kraft inequality, we can construct an instantaneous code by Theorem 5.2.1. Every instantenous code is uniquely decodable, so we are done. □

5.3 Optimal Codes

After we have the bound, how do we actually find the instantaneous code with minimum expected length? This is equivalent to the optimization problem

$$\begin{aligned} &\text{Minimize} && L = \sum_{i=1}^{m} p_i\ell_i \\ &\text{with constraint} && \sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1, \quad \ell_i \in \mathbb{Z}, \forall i \end{aligned} \tag{OPT}$$

Theorem 5.3.1: Expected Length Bound

The expected code length L of any instantaneous binary code for a r.v. X satisfies

$$L \geqslant H(X)$$

with equality if and only if $p_i = 2^{-\ell_i}$.

Proof 1. We consider the optimization problem OPT and ignore the integer constraint. Using Lagrangian multiplier,

$$J = \sum_{i=1}^{m} p_i\ell_i + \lambda\left(\sum_{i=1}^{m} 2^{-\ell_i} - 1\right)$$

Differentiate w.r.t. ℓ_i and take it to 0, we have

$$\frac{\partial J}{\partial \ell_i} = p_i - \lambda 2^{-\ell_i}\log_e 2 = 0 \quad\Longrightarrow\quad 2^{-\ell_i} = \frac{p_i}{\lambda\log_e 2}$$

Since the constraint is in active, we substitute this in $\sum 2^{-\ell_i} = 1$, and get

$$\sum_{i=1}^{m}\frac{p_i}{\lambda\log_e 2} = \frac{1}{\lambda\log_e 2}\sum_{i=1}^{m} p_i = \frac{1}{\lambda\log_e 2} = 1 \quad\Longrightarrow\quad \lambda = \frac{1}{\log_e 2}$$

which yields $p_i = 2^{-\ell_i}$ and the optimal code lengths $\ell_i^\star = -\log_2 p_i$. Then, the expected length is

$$L^\star = \sum_{i=1}^{m} p_i\ell_i^\star = -\sum_{i=1}^{m} p_i\log_2 p_i = H(X)$$

But since ℓ_i must be integers, we will not always be able to get this bound. □

Proof 2. We write

$$L - H(X) = \sum_{i=1}^{m} p_i \ell_i - \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{m} p_i \log_2 2^{-\ell_i} + \sum_{i=1}^{m} p_i \log p_i$$

Let $r_i = 2^{-\ell_i} / \sum_j 2^{-\ell_j}$ and $c = \sum_i 2^{-\ell_i}$, we have

$$L - H(X) = \sum_{i=1}^{m} p_i \log \frac{p_i}{r_i} - \log c = D(\mathbf{p} \| \mathbf{r}) + \log \frac{1}{c} \geqslant 0$$

where the last step follows from that relative entropy is non-negative, and Kraft inequality provides that $c \leqslant 1$. Hence, $L \geqslant H(X)$ with equality if and only if $\mathbf{p} = \mathbf{r}$, and $c = 1$. This happens if and only if $p_i = 2^{-\ell_i}$. □

However, if we want to choose the optimal code, we need to find the distribution $p_i = 2^{-\ell_i}$ that is closest to distribution of X, which is not obvious. Then, how do we choose the optimal code?

5.4 Suboptimal Procedure: Shannon-Fano Coding

In the last section, we show that $\ell_i = \log_2 \frac{1}{p_i}$ yields the optimal code. Since this may not be an integer, we round it up to $\ell_i = \lceil \log_2 \frac{1}{p_i} \rceil$. These lengths also satisfy Kraft inequality since $\sum 2^{-\lceil \log_2 \frac{1}{p_i} \rceil} \leqslant \sum 2^{\log_2 \frac{1}{p_i}} = \sum p_i = 1$. This choice of codeword satisfies $\log_2 \frac{1}{p_i} \leqslant \ell_i < \log_2 \frac{1}{p_i} + 1$. Therefore, multiply by p_i and summing over i, we obtain the bound on the optimal code length:

$$H(X) \leqslant L < H(X) + 1$$

This is an overhead at most 1 bit. However, we can reduce the overhead per symbol by spreading it out over many symbols.

i.i.d. Case: Suppose we observe a sequence $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$, where $(x_1, \cdots, x_n) \in \mathcal{X}^n$. If $\ell(x_1, \cdots, x_n)$ denotes the length of the binary codeword associated with $(x_1, \cdots, x_n)$, then we define L_n to be the expected codeword length per input symbol, that is

$$L_n = \frac{1}{n} \sum_{x_1, \cdots, x_n} p(x_1, \cdots, x_n) \ell(x_1, \cdots, x_n) = \frac{1}{n} \mathbb{E}\left[\ell(X_1, \cdots, X_n)\right]$$

Apply the bound above,

$$H(X_1, \cdots, X_n) \leqslant \mathbb{E}[\ell(X_1. \cdots, X_n)] < H(X_1, \cdots, X_n) + 1$$

Since $X_1, \cdots, X_n$ are i.i.d., $H(X_1, \cdots, X_n) = \sum H(X_i) = nH(X)$. Therefore, divide by n on the inequality above, we have that

$$H(X) \leqslant L_n < H(X) + \frac{1}{n}$$

Therefore, by using large block lengths we can achieve an expected codelength per symbol arbitrarily close to the entropy.

Stochastic Process Case: This can be achieved also for stochastic process. In this case, we still have

$$\frac{H(X_1, \cdots, X_n)}{n} \leqslant L_n < \frac{H(X_1, \cdots, X_n)}{n} + \frac{1}{n}$$

If the stochatic process is *stationary*, then $H(X_1, \cdots, X_n)/n \to H(\mathbb{X})$, and the expected description length tends to the entropy rate as $n \to \infty$.

Finally, if we have the wrong pmf estimation $q(x) \neq p(x)$, we will not achieve expected length $L \approx H(p)$.

Theorem 5.4.1: Suboptimality of Shannon-Fano

The expected length under $p(x)$ of the code assignment $\ell(x) = \lceil \log \frac{1}{q(x)} \rceil$ satisfies

$$H(p) + D(p\|q) \leqslant \mathbb{E}_p[\ell(X)] < H(p) + D(p\|q) + 1$$

Proof. For lower bound,

$$\mathbb{E}[\ell(X)] = \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil \geqslant \sum_x p(x) \log \frac{1}{q(x)} = \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} = H(p) + D(p\|q)$$

For the upper bound,

$$\mathbb{E}[\ell(X)] = \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil \leqslant \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) = \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 = H(p) + D(p\|q) + 1$$

which completes the proof. □

Remark: This section justifies the definition of entropy and relative entropy:

- **Entropy rate** is the expected number of bits per symbol required to describe the process.
- Believing that the distribution is $q(x)$ when true distribution is $p(x)$ incurs a penalty of **relative entropy** $D(p\|q)$ in the average description length.

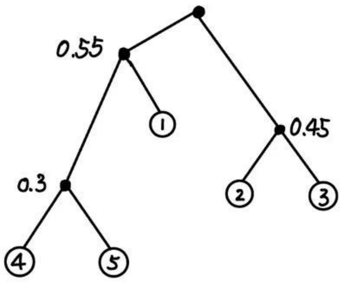
5.5 Optimal Procedure: Huffman Coding

Huffman Codes:

- Input $p_1, \cdots, p_n$.
- Choose the two least probable symbols in the alphabet.
- Delete the two symbols, and add a new symbol as the parent node of the previous two symbols in the binary tree.

Example 1: Consider a random variable X taking values in $\mathcal{X} = \{1, 2, 3, 4, 5\}$ with probabilities, $0.25, 0.25, 0.2, 0.15, 0.15$, respectively. The algorithm goes as follows:

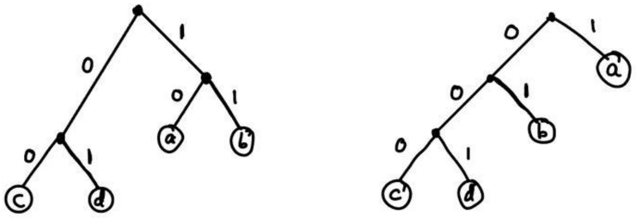
$$X = \begin{cases} 1, & \text{w.p. } 0.25, \\ 2, & \text{w.p. } 0.25, \\ 3, & \text{w.p. } 0.2, \\ 4, & \text{w.p. } 0.15, \\ 5, & \text{w.p. } 0.15 \end{cases} \implies \begin{cases} 1, & \text{w.p. } 0.25, \\ 2, & \text{w.p. } 0.25, \\ 3, & \text{w.p. } 0.2, \\ (4,5), & \text{w.p. } 0.3 \end{cases} \implies \begin{cases} 1, & \text{w.p. } 0.25, \\ (2,3), & \text{w.p. } 0.45, \\ (4,5), & \text{w.p. } 0.3 \end{cases} \implies \begin{cases} (1,4,5), & \text{w.p. } 0.55, \\ (2,3), & \text{w.p. } 0.45 \end{cases}$$



Example 2: Consider a random variable X taking values in $\mathcal{X} = \{a, b, c, d\}$ with probabilities $1/3, 1/3, 1/4, 1/12$, respectively. We demonstrate two ways of Huffman coding.

$$X = \begin{cases} a, & \text{w.p. } 1/3, \\ b, & \text{w.p. } 1/3, \\ c, & \text{w.p. } 1/4, \\ d, & \text{w.p. } 1/12 \end{cases} \implies \begin{cases} a, & \text{w.p. } 1/3, \\ b, & \text{w.p. } 1/3, \\ (c,d), & \text{w.p. } 1/3 \end{cases} \implies \begin{cases} (a,b), & \text{w.p. } 2/3, \\ (c,d), & \text{w.p. } 1/3 \end{cases}$$

$$X = \begin{cases} a, & \text{w.p. } 1/3, \\ b, & \text{w.p. } 1/3, \\ c, & \text{w.p. } 1/4, \\ d, & \text{w.p. } 1/12 \end{cases} \implies \begin{cases} a, & \text{w.p. } 1/3, \\ b, & \text{w.p. } 1/3, \\ (c,d), & \text{w.p. } 1/3 \end{cases} \implies \begin{cases} a, & \text{w.p. } 1/3, \\ (b,c,d), & \text{w.p. } 2/3 \end{cases}$$



Remark: Using Shannon coding from last section, we have codeword lengths of $\lceil \log \frac{1}{p_i} \rceil$, may be much worse than the optimal code. For example, consider two symbols, one with probability 0.9999 and the other one with probability 0.0001. Then, using Shannon code gives codeword length of $\lceil \log \frac{1}{0.0001} \rceil = 14$ bits and $\lceil \log \frac{1}{0.9999} \rceil = 1$ bit. The optimal is obvious 1 bit for both symbols.

Now we examine the optimality of Huffman code.

Lemma 5.5.1: Canonical Code

For any distribution, there exists an optimal prefix code s.t. $p_1 \geqslant p_2 \geqslant \cdots p_m$ with $\ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_{m-1} = \ell_m$, and $C(x_m)$, $C(x_{m-1})$ only differs in last bit. This code is called *canonical*.

Proof. The existence of such code is demonstrated by the Huffman coding algorithm. Now we need to show that it is optimal.
Let $p = (p_1, p_2, \cdots, p_n)$ with $p_1 \geqslant p_2 \geqslant p_n$ be some distribution. Define the Huffman reduction $p' = (p_1, \cdots, p_{m-1}+p_m)$. Let $C^\star_{m-1}(p')$ be optimal code for p', and let $C^\star_m(p)$ be optimal canonical code for p.
PART I: We expand $C^\star_{m-1}(p')$ to a code for p. This is obtained by the following algorithm: take the codeword corresponding to weight $p_{m-1}+p_m$ and adding a 0 to form a codeword for symbol $m-1$ and adding 1 to form a codeword for symbol m.

$$\begin{array}{ccc} & C^\star_{m-1}(p') & \\ \text{distribution} & \text{codeword} & \text{length} \\ p_1 & w'_1 & \ell'_1 \\ p_2 & w'_2 & \ell'_2 \\ \vdots & \vdots & \vdots \\ p_{m-2} & w'_{m-2} & \ell'_{m-2} \\ p_{m-1}+p_m & w'_{m-1} & \ell'_{m-1} \end{array} \Longrightarrow \begin{array}{ccc} & C_m(p) & \\ \text{distribution} & \text{codeword} & \text{length} \\ p_1 & w'_1 & \ell'_1 \\ p_2 & w'_2 & \ell'_2 \\ \vdots & \vdots & \vdots \\ p_{m-2} & w'_{m-2} & \ell'_{m-2} \\ p_{m-1} & w'_{m-1}0 & \ell'_{m-1}+1 \\ p_m & w'_{m-1}1 & \ell'_{m-1}+1 \end{array}$$

We have the expected length

$$L(p) = L^\star(p') + p_{m-1} + p_m \tag{5.2}$$

PART II: We condense an optimal canonical code for p to construct a code for the Huffman reduction p'. This is done by merging the codewords for the two lowest probability symobls $m-1$ and m.

$$\begin{array}{ccc} & C^\star_m(p) & \\ \text{distribution} & \text{codeword} & \text{length} \\ p_1 & w_1 & \ell_1 \\ p_2 & w_2 & \ell_2 \\ \vdots & \vdots & \vdots \\ p_{m-1} & w_{m-1} & \ell_{m-1} \\ p_m & w_m & \ell_m \end{array} \Longrightarrow \begin{array}{ccc} & C_{m-1}(p') & \\ \text{distribution} & \text{codeword} & \text{length} \\ p_1 & w_1 & \ell_1 \\ p_2 & w_2 & \ell_2 \\ \vdots & \vdots & \vdots \\ p_{m-1}+p_m & \bar{w}_{m-1} = \bar{w}_m & \ell_{m-1}-1 = \ell_m - 1 \end{array}$$

where $\bar{w}_m$ denotes the codeword w_m without last bit, and since $\ell_{m-1} = \ell_m$, we have the equation indicated above. Thus, the code expected length is

$$L(p') = \sum_{i=1}^{m-2} p_i\ell_i + (p_{m-1}+p_m)(\ell_{m-1}-1) = \sum_{i=1}^{m-2} p_i\ell_i + p_{m-1}(\ell_{m-1}-1) + p_m(\ell_m - 1) = L^\star(p) - p_{m-1} - p_m \tag{5.3}$$

Combining equation 5.2 and 5.3 together, we have

$$(L(p') - L^\star(p')) + (L(p) - L^\star(p)) = 0$$

Since $L^\star(p')$ and $L^\star(p)$ are optimal, we have both $(L(p') - L^\star(p')) \geqslant 0$ and $(L(p) - L^\star(p)) \geqslant 0$. Therefore, the only way to make the summation to zero is to let $L(p') = L^\star(p')$ and $L(p) = L^\star(p)$. Consequently, the code is optimal. □

Thus, Huffman code is optimal.

Theorem 5.5.2: Optimality of Huffman Coding

If $C^\star$ is a Huffman code and C' is any other uniquely decodable code, then $L(C^\star) \leqslant L(C')$.

5.6 Generating Discrete Distributions from Fair Coins

Question considered so far: Compressing a r.v. X into sequence of bits, and minimizing number of bits.
Dual question: How many fair coin flips does it take to generate a random variable X drawn according to a specified pmf p?
Example 1: Generate X s.t. $X = a$ w.p. $1/2$, $X = b$ w.p. $1/4$ and $X = c$ w.p. $1/4$.
Answer: If first flip is H, let $X = a$, if first flip is T, then second flip H, $X = b$, second flip T, $X = c$. Average number of flips $= 1.5$ bits $=$ entropy of the distribution.

General Problem: Given sequence of fair coin tosses $Z_1, Z_2, \cdots$, generate a r.v. $X \in \mathcal{X} = \{1, 2, \cdots, m\}$ with pmf $p = (p_1, \cdots, p_m)$. Let r.v. T denote the number of coin flips in the algorithm. An algorithm mpas bits $Z_1, Z_2, \cdots$ to possible outcomes X can be described by a binary tree.

- The tree should be complete (may be infinite).
- The probability of a leaf at depth k is 2^{-k}. (Many leaves may be labeled with the same output symbol).
- The expected number of fair bits $\mathbb{E}[T]$ required to generate X is equal to the expected depth of the tree.

What is the most efficient algorithm to generate a given distribution?

Lemma 5.6.1: Expected depth of the tree

Let $\mathcal{Y}$ denote the set of leaves of a complete tree. Consider the distribution of the tree s.t. the probability of a leaf of depth k is 2^{-k}. Let Y be a r.v. with this distribution, then

$$\mathbb{E}[T] = \sum_{y \in \mathcal{Y}} k(y) 2^{-k(y)} = H(Y)$$

Proof.

$$H(Y) = -\sum_{y \in \mathcal{Y}} \frac{1}{2^{k(y)}} \log \frac{1}{2^{k(y)}} = \sum_{y \in \mathcal{Y}} k(y) 2^{-k(y)} = \mathbb{E}[T]$$

□

Theorem 5.6.2: Expected Fair Bits

For any algorithm generating X, the expected number of fair bits used is greater than the entropy $H(X)$, i.e.,

$$\mathbb{E}[T] \geqslant H(X)$$

Proof. Any algorithm generating X from fair bits can be represented by a compelte binary tree. Label all the leaves of this tree by distinct symbols $y \in \mathcal{Y} = \{1, 2, \cdots\}$. Let Y denote the distribution on the tree as in the Lemma 5.6.1. By lemma, $\mathbb{E}[T] = H(Y)$. Now, the r.v. X is a function of Y (one or more leaves map onto an output symbol), hence $H(X) \leqslant H(Y)$ by Proposition 2.1.6, which complete the proof. □

Theorem 5.6.3: Dyadic Distribution Achieves the Bound

Let the r.v. X satisfy that $\log \frac{1}{p(x)}$ is an integer for all $x \in \mathcal{X}$, then there exists an algorithm to generate X from fair coin flips with $\mathbb{E}[T] = H(X)$.

Proof. For the upper bound, use Huffman code. When the distribution is dyadic, the Huffman code is the same as the Shannon code, and achieves the entropy bound, with depth of a leaf in the tree corresponding to x is $\log \frac{1}{p(x)}$. Using this tree, the leaf will have probability $2^{-\log \frac{1}{p(x)}} = p(x)$. Thus, the expected number of coin flips is the expected depth of the tree, which is the achieved entropy bound. □

Indeed, the expected number of flips is bounded by $H(X) + 2$, see Thomas and Joy[1] p137-141.

Bibliography

[1] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

[2] Polyanskiy, Y. and Wu, Y. (2024). *Information theory: From coding to learning*. Cambridge university press.