Contents

1 Probability

1.1 Probability Spaces

In an elementary level, we have been viewing probability as the quotient between the number of desired outcomes and the number of all possible outcomes. This definition, though intuitive, is not very solid when it comes to an infinite sample space. In this introductory chapter, we would establish the theories of probability using a more modern and rigorous structure.

Definition 1.1.1 ▶ Set Algebra

Let X be a set. A **set algebra** over X is a family $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

- $X \backslash F \in \mathcal{F}$ for all $F \in \mathcal{F}$ (closed under complementation);
- $X \in \mathcal{F}$;
- $X_1 \cup X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$ (closed under binary union).

There are several immediate implications from the above definition.

First, by closure under complementation, we know that an algebra over any set X must contain the empty set.

Second, by De Morgan's Law, one can easily check that if the first 2 axioms hold, the closure under binary union is equivalent to

- $X_1 \cap X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$;
- $\bigcup_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$;
- $\bigcap_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

$(X, \mathcal{F})$ is known as a *field of sets*, where the elements of X are called *points* and those of $\mathcal{F}$, *complexes* or *admissible sets* of X.

In probability theory, what we are interested in is a special type of set algebras known as *σ-algebras*.

Definition 1.1.2 ▸ σ-Algebra

A **σ-Algebra** over a set A is a non-empty set algebra over A that is closed under countable union.

Of course, by the same argument as above, we known that any σ-algebra is closed under countable intersection as well.

Now, as we all know, we can take some set Ω as a *sample space* and denote an *event* by some subset of Ω. Roughly speaking, we could now define the probability of an event $E \subseteq \Omega$ as the ratio between the sets' volumes. The remaining question now is: how do we define the volume of a set properly?

Definition 1.1.3 ▸ Measure

Let X be a set and Σ be a σ-algebra over X. A **measure** over Σ is a function

$$\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$$

such that

- $\mu(E) \geq 0$ for all $E \in \Sigma$ (non-negativity);
- $\mu(\varnothing) = 0$;
- $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$ for any countable collection of pairwise disjoint elements of Σ (countable additivity or σ-additivity).

The triple (X, Σ, μ) is known as a **measure space** and the pair (X, Σ), a **measurable space**.

One thing to note here is that if at least one $E \in \Sigma$ has a finite measure, then $\mu(\varnothing) = 0$ is automatically guaranteed for obvious reasons.

Definition 1.1.4 ▸ Probability Space

Let Ω be a sample space and $\mathcal{F}$ be a σ-algebra over Ω. A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{P} : \mathcal{F} \to [0, 1]$, known as a **probability measure**, is such that $\mathbb{P}(\Omega) = 1$.

Obviously, the above definition immediately guarantees that

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
2. $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $\mathbb{P}(A) \subseteq \mathbb{P}(A)$;
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

The third result follows from a direct application of the principle of inclusion and exclusion.

By induction, one can easily check that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(E_i\right)$$

for any finitely many events. The following proposition extends this result to countable collections of events:

Proposition 1.1.5 ▸ Union Bound of Countable Collections of Events

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E_1, E_2, \cdots, E_n, \cdots \in \mathcal{F}$ is any countable sequence of events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}\left(E_i\right).$$

Proof. Define $F_1 \coloneqq E_1$ and $F_k \coloneqq E_k \backslash \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Clearly, the F_i's are pairwise disjoint. By Definition 1.1.2, the F_i's are elements of $\mathcal{F}$. Note that $\mathbb{P}\left(F_i\right) \leq \mathbb{E}_{\mathbb{i}}$ for all $i \in \mathbb{N}^+$, so

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}\left(F_i\right) \\
&\leq \sum_{i=1}^{\infty} \mathbb{P}\left(E_i\right).
\end{aligned}$$

□

Next, we will introduce the notion of *random variables* formally. For this purpose, we first establish the notion of a *Borel algebra*.

Definition 1.1.6 ▸ Borel Algebra

Let X be a topological space. A **Borel set** on X is a set which can be formed via countable union, countable intersection and relative complementation of open sets in X. The smallest σ-algebra over X containing all Borel sets on X is known as the **Borel algebra** over X.

Clearly, the Borel algebra over X contains all open sets in X according to the above axioms from Definition 1.1.2. This helps us define the following:

Definition 1.1.7 ▸ Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is the Borel algebra over $\mathcal{X}$. A **random variable** is a function $X : \Omega \to \mathcal{X}$ such that

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for all $B \in \mathcal{B}$.

Remark. Rigorously, such a random variable X is a *measurable function* or *measurable mapping* from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

The probability measure $\mathbb{P}$ thus induces a probability measure P_X over $(\mathcal{X}, \mathcal{B})$.

Definition 1.1.8 ▸ Distribution

Let $X : \Omega \to \mathcal{X}$ be a random variable over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{B}$ be the Borel algebra over $\mathcal{X}$, the **distribution** of X is the probability measure P_X on $(\mathcal{X}, \mathcal{B})$ given by

$$P_X(B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

Remark. Often times, we write $\Pr(X \in B) = P_X(B)$.

In the context of information theory, we mostly are concerned with real-valued random variables only.

Definition 1.1.9 ▸ Real-Valued Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, a **real-valued random variable** over the space is a mapping $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$.

Note that the Borel set over $\mathbb{R}$ is just the family of all open intervals.

Clearly, if X is a real-valued random variable, we have $\{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}$. Moreover, we claim that

$$\{\omega \in \Omega : X(\omega) < x\} = \bigcup_{y<x} \{\omega \in \Omega : X(\omega) \leq y\}.$$

The proof is quite straightforward and is left to the reader as an exercise. By Definition

1.1.2, this means that

$$\{\omega \in \Omega : X(\omega) < x\} \cup \{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}.$$

Therefore, $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$. This argument justifies the probabilities $\Pr(X < x)$ and $\Pr(X = x)$. We give a special name to the range of a random variable in computer science.

Definition 1.1.10 ▸ Alphabet

Let X be a random variable, the range of X is called an **alphabet**, denoted as $\mathcal{X}$.

Recall that we have defined expectations for discrete and continuous random variables in elementary probability theory. In terms of measure theory, the two formulae can be unified as the Lebesgue integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbb{P}(\omega).$$

Note that $\mathbb{E}[X]$ is a real number while $\mathbb{E}[X \mid Y]$ is a **random variable** formed as a function of Y. In a way, Y partitions the sample space into regions where $\mathbb{E}[X \mid Y = y_i]$ gives the expectation of X in the region induced by $Y = y_i$ for each $y_i \in \mathcal{Y}$. In general, the following result holds:

Theorem 1.1.11 ▸ Law of Iterated Expectations

Let X and Y be random variables, then $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$.

The above formula can be interpreted as the fact that $\mathbb{E}[X \mid Y]$ is a best estimator for X.

1.2 Markov Chains

Recall that 2 random variables X and Z are *independent* if and only if $P_{X,Z}(x,z) = P_X(x) P_Z(z)$ or $P_{X|Z}(x \mid z) = P_X(x)$ for all $(x,z) \in \mathcal{X} \times \mathcal{Z}$. We will extend this definition with a third random variable.

Definition 1.2.1 ▸ Conditional Independence

Let X, Y, Z be random variables. If

$$P_{X,Y,Z}(x,y,z) = P_X(x) P_{Y|X}(y \mid x) P_{Z|Y}(z \mid y)$$

for all $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then we say that X, Y, Z forms a **Markov chain** in this order, or that X and Z are **conditionally independent** on Y.

Recall also that the *Bayes's Rule* states the following:

Theorem 1.2.2 ▶ Bayes's Rule

For any random variables X and Y,

$$P_{X|Y}(x \mid y) = \frac{P_{Y|X}(y \mid x) P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y \mid x') P_X(x')}.$$

Based on Theorem 1.2.2, we have

$$P_{X,Y}(x,y) = P_{X|Y}(x \mid y) P_Y(y) = P_X(x) P_{Y|X}(y \mid x).$$

By applying the formula repeatedly, we have

$$\begin{aligned} P_{X,Y,Z}(x,y,z) &= P_{X,Y}(x,y) P_{Z|X,Y}(z \mid x,y) \\ &= P_X(x) P_{Y|X}(y \mid x) P_{Z|X,Y}(z \mid x,y). \end{aligned}$$

Therefore, a Markov chain simply states that the distribution of Z is no longer dependent on X, but depends on Y solely. Therefore, this allows us to remove one condition when applying Theorem 1.2.2. Thus, it actually suffices to prove $P_{Z|X,Y} = P_{Z|Y}$ when proving that X-Y-Z forms a Markov chain.

We can denote a Markov chain by X-Y-Z. Intuitively, such a relationship should be symmetric.

Proposition 1.2.3 ▶ Symmetricity of Markov Chains

If X-Y-Z is a Markov chain, then Z-Y-X is also a Markov chain.

Proof. By Definition 1.2.1,

$$P_{X,Y,Z}(x,y,z) = P_X(x) P_{Y|X}(y \mid x) P_{Z|Y}(z \mid y).$$

By Theorem 1.2.2, we have

$$\begin{aligned} P_{X|Y}(x \mid y) &= \frac{P_X(x) P_{Y|X}(y \mid x)}{P_Y(y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_Y(y) P_{Z|Y}(z \mid y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_{Z,Y}(z,y)} \\ &= P_{X|Z,Y}(x \mid z,y). \end{aligned}$$

Therefore, Z-Y-X is a Markov chain. □

One obvious case where dependence exists between the random variables in a Markov chain is that one of the random variables is a function of another one.

Proposition 1.2.4 ▸ Markov Chain Involving Functions of a Random Variable

Let X and Y be any random variables and $Z \coloneqq f(Y)$ for some function f, then X-Y-Z is a Markov chain.

Proof. Notice that

$$P_{Z|X,Y}(z \mid x, y) = P_{f(Y)|X,Y}(z \mid x, y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases},$$

$$P_{Z|Y}(z \mid y) = P_{f(Y)|Y}(z \mid y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases}$$

for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Therefore, $P_{Z|X,Y} = P_{Z|Y}$ and so X-Y-Z forms a Markov chain. □

Note that if X and Z are independent, they are naturally conditionally independent given any Y. However, the inverse may not be true.

Proposition 1.2.5 ▸ Conditional Independence Does Not Imply Independence

There exists random variables X, Y, Z such that X and Z are dependent but conditionally independent given Y.

Proof. Let N_1, N_2, N_3 be pairwise independent random variables such that

$$\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_3 = \{0, 1\}.$$

Take $X = N_1 + N_2$, $Y = N_2$ and $Z = N_2 + N_3$. Clearly, X and Z are dependent, but

$$\begin{aligned} P_{Z|X}(z \mid x) &= P_{N_2+N_3|N_1+N_2}(z \mid x) \\ &= P_{N_3|N_1,N_2}(z - y \mid x - y, y) \\ &= P_{N_2+N_3|N_1+N_2,N_2}(z \mid x, y) \\ &= P_{Z|X,Y}(z \mid x, y), \end{aligned}$$

which implies that X and Z are conditionally independent given Y. □

1.3 Probability Bounds

We use various bounds to make estimates and approximations for probability distributions. The first commonly used bound is *Markov's Inequality*.

Theorem 1.3.1 ▸ Markov's Inequality

If X is a non-negative random variable, then $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ *for all* $a > 0$.

Proof. It suffices to prove for the continuous case. Notice that

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f_X(x)\,\mathrm{d}x \\ &\geq \int_a^\infty x f_X(x)\,\mathrm{d}x \\ &\geq a\int_0^\infty f_X(x)\,\mathrm{d}x \\ &= \Pr(X \geq a).\end{aligned}$$

Therefore, $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. □

Note that the bound given by Markov's inequality is a rather loose bound. The following inequality proposes a better bound:

Theorem 1.3.2 ▸ Chebyshev's Inequality

For any real-valued random variable X with finite variance,

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{a^2}$$

for all $a > 0$.

Proof. Define $g(X) : (X - \mathbb{E}[X])^2$, which is clearly non-negative. By Theorem 1.3.1, we have

$$\Pr\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{\mathbb{E}[g(X)]}{a^2\mathrm{Var}(X)}.$$

Note that $\mathbb{E}[g(X)] = \mathrm{Var}(X)$, so

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) = \Pr\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{1}{a^2}.$$

□

Finally, we state the following law of large numbers:

Theorem 1.3.3 ▶ Weak Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$. For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

Proof. Note that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$ and that

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sum_{i=1}^{n} \operatorname{Var}(X_i)}{n^2} = \frac{\sigma^2}{n}.$$

By Theorem 1.3.2, we have

$$0 \leq \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

By Squeeze Theorem, this clearly implies that

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

□

Alternatively, we may phrase Theorem 1.3.3 as "$\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ in probability". When a sequence $\{S_n\}_{n=1}^{\infty}$ converges to b in probability, we write $S_n \xrightarrow{\text{p}} b$.

> *Remark.* Essentially, what Theorem 1.3.3 says is that when n is large, the sample mean from n measurements of the same data converges to the expectation of the distribution.

Under some mild conditions, this convergence occurs exponentially fast, i.e., the probability $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right)$ decreases at least as fast as $\exp(-ng(\epsilon))$ for some real-valued function $g : \mathbb{R}^+ \to \mathbb{R}^+$. In terms of asymptotic analysis, we write this as

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \exp(-ng(\epsilon) + o(n)).$$

Equivalently, this means that there exists a function $g : \mathbb{R} \to \mathbb{R}$ with $g(\epsilon) > 0$ for every

$\epsilon > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \geq g(\epsilon) + o(1).$$

There is a strong version for the law, which shall be stated without proof:

Theorem 1.3.4 ▸ Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$, then

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

1.4 Convexity

Recall the following definition:

Definition 1.4.1 ▸ Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for any $\lambda \in [0,1]$ and any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

From a graphical perspective, a convex function is an overestimate of all linear functions whose values are bounded above by it. The following proposition set this result in a rigorous context:

Proposition 1.4.2 ▸ Convex Functions as Overestimates for Linear Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and define

$$\mathcal{L} \coloneqq \left\{\ell \in \operatorname{Maps}(\mathbb{R}^n, \mathbb{R}) : \ell(\boldsymbol{u}) = \boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{u} + b \leq f(\boldsymbol{u}) \text{ for all } \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}\right\}$$

to be the set of all linear functions bounded above by f, then for each $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) = \sup_{\ell \in \mathcal{L}} \ell(\boldsymbol{x}).$$

Proof. It suffices to prove that for all $\boldsymbol{x} \in \mathbb{R}^n$, there exists some linear function $\ell \in \mathcal{L}$

such that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take any $\boldsymbol{h} \in \mathbb{R}^n$. Since f is convex, we have

$$\begin{aligned} 2f(\boldsymbol{x}) &= 2f\left(\frac{1}{2}(\boldsymbol{x}+\boldsymbol{h}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{h})\right) \\ &\leq f(\boldsymbol{x}+\boldsymbol{h}) + f(\boldsymbol{x}-\boldsymbol{h}). \end{aligned}$$

Therefore, we have

$$L_1 = \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}-\boldsymbol{h})}{\|\boldsymbol{h}\|} \leq \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|} = L_2.$$

Take some $a \in [L_1, L_2]$ and let $\ell(\boldsymbol{y}) = a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x})$. Observe that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take $\boldsymbol{h} = \boldsymbol{y} - \boldsymbol{x}$, then

$$\begin{aligned} \ell(\boldsymbol{y}) &= a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &\leq \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|}\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &= f(\boldsymbol{x}+\boldsymbol{h}) \\ &= f(\boldsymbol{y}). \end{aligned}$$

Therefore, $\ell \in \mathcal{L}$ as desired. □

The following proposition gives a simple test for convexity in one-dimensional case, which is a special case of the Hessian matrix test:

Proposition 1.4.3 ▸ Second Derivative Test for Convexity

If a real-valued function f is twice-differentiable on $[a, b]$, then it is convex if and only if $f''(x) \geq 0$ for all $x \in (a, b)$.

Convex functions produce the following interesting result regarding expectation:

Theorem 1.4.4 ▸ Jensen's Inequality

Let f be a convex function and X be a random variable, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Proof. Let $\mathcal{L}$ be the set of all linear functions bounded above by f. By Proposition

1.4.2, we have

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[\sup_{\ell\in\mathcal{L}} \ell(X)\right] \\ &\geq \sup_{\ell\in\mathcal{L}} \mathbb{E}[\ell(X)] \\ &= \sup_{\ell\in\mathcal{L}} \ell(\mathbb{E}[X]) \\ &= f(\mathbb{E}[X]).\end{aligned}$$

□

Remark. If f is strictly convex, the equality holds if and only if X is constant.

2

Information Theory

2.1 Entropy

In information theory, the very first question to ask is how we can measure the quantity of information contained in communication. Colloquially, we say that communication gives more information if more knowledge which has remained unknown previously is revealed.

We describe such revelation of new knowledge as the "surprise" of an event. Using probability theory, we use a random variable X to represent an event by $X = x$. Intuitively, an event is surprising if the probability of its occurrence is low. This is formally stated as follows:

Definition 2.1.1 ▸ Surprise

Let X be a random variable. The **surprise** of an event $X = x$ is defined as

$$\log_2 \frac{1}{P_X(x)} = \log_2 \frac{1}{\Pr(X = x)}.$$

Now, suppose we are **uncertain** about some event $X = x$. We may wish to measure how much uncertainty we have towards the outcome of the event, or equivalently, what is the **expected surprise** for the event. It is easy to see that if we define a random variable for surprise as a function of X, we can make use of the expectation formula to compute this quantity.

Definition 2.1.2 ▸ Entropy of Discrete Random Variables

Let X be a discrete random variable supported on a finite alphabet $\mathcal{X}$ with probability mass function p_X, then **entropy** of X is defined as

$$H(X) \coloneqq -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x).$$

It is clear that this definition can be manipulated into

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{p_X(X)}\right],$$

i.e., the entropy of X is exactly the expected surprise of X. There is a small problem, though,

which is that $\log_2 n$ is undefined when $n \leq 0$. Since $p_X(x) \geq 0$ for all $x \in \mathcal{X}$, we only need to take care of 0 as a special case. Notice that

$$\begin{aligned}\lim_{x\to 0^+} x \log_2 x &= \lim_{x\to 0^+} \frac{x \ln x}{\ln 2} \\ &= -\frac{1}{\ln 2} \lim_{x\to 0^+} \frac{-\ln x}{x^{-1}} \\ &= -\frac{1}{\ln 2} \lim_{x\to 0^+} \frac{-x^{-1}}{-x^{-2}} \\ &= 0.\end{aligned}$$

Therefore, it makes sense to set $x \log_2 x = 0$ when $x = 0$.

> *Remark.* By convention, we set $0 \log_2 0 = 0$.

We will later prove that $0 \leq H(X) \leq \log_2 |\mathcal{X}|$. Moreover, $H(X)$ is closely related with the minimal number of bits to encode X in binary number unambiguously. In particular, if we let $b(X)$ be the minimal number of bits to encode X in binary strings unambiguously, we have

$$H(X) \leq \mathbb{E}[b(X)] < H(X) + 1.$$

Moreover, if we let $q(X)$ to be the number of attempts to guess the value of X correctly, we might be surprised by the fact that

$$H(X) \leq \mathbb{E}[q(X)] < H(X) + 1,$$

i.e., it is expected to attempt at least $H(X)$ times to guess the value of X, but there is always a strategy to expect success before the $\big(H(X)+1\big)$-th attempt.

Those with prior knowledge in machine learning and decision trees might find the following special form of entropy familiar:

Definition 2.1.3 ▸ Binary Entropy

Let X be a Bernoulli random variable with parameter p. The **binary entropy** of p is defined as

$$H_b(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

With some simple computation, it is easy to check that $H_b(p)$ is maximised when $p = \frac{1}{2}$ and is zero if and only if $p = 1$ or $p = 0$.

Entropy can be defined over multiple random variables just like probability distributions.

In fact, we denote the tuple of n random variables as

$$X_1^n := (X_1, X_2, \cdots, X_n)$$

Clearly, we may view X_1^n as nothing else than a single random variable whose alphabet is just $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$.

Definition 2.1.4 ▸ Joint Entropy

Let X_1^n be a tuple of discrete random variable supported on a finite alphabet

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$$

with joint probability mass function $p_{X_1^n}$. The **joint entropy** of $X_1, X_2, \cdots, X_n$ is defined as

$$H(X_1^n) := -\sum_{\boldsymbol{x} \in \mathcal{X}} p_{X_1^n}(\boldsymbol{x}) \log_2 p_{X_1^n}(\boldsymbol{x}).$$

Additionally, we can of course define the conditional entropy to measure the uncertainty of one event given the information on another event.

Definition 2.1.5 ▸ Conditional Entropy

Let (X, Y) be a pair of discrete random variables supported on an alphabet $\mathcal{X} \times \mathcal{Y}$ which is finite. Let p_X and p_Y be the probability mass functions for X and Y respectively. The **conditional entropy** of X given Y is defined as

$$H(X \mid Y) := \sum_{y \in \mathcal{Y}} p_Y(y) H(X \mid Y = y).$$

Note that here $H(X \mid Y = y)$ is also known as the *conditional entropy*, but it is different in meaning with $H(X \mid Y)$. In particular:

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 p_{X|Y}(x \mid y).$$

Therefore, we can expand the expression in Definition 2.1.5 into

$$\begin{aligned} H(X \mid Y) &= -\sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 p_{X|Y}(x \mid y) \\ &= -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x, y) \log_2 p_{X|Y}(x \mid y) \\ &= \mathbb{E}\left[\log_2 \frac{1}{p_{X|Y}(X \mid Y)}\right]. \end{aligned}$$

One thing to note is that conditional entropy is **not symmetric**. We can interpret $H(X \mid Y)$ as "the remaining uncertainty of X given information on Y". Hence, it is not surprising that the following identity is true:

$$H(X,Y) = H(X) + H(Y \mid X)$$

This is generalised as follows:

Proposition 2.1.6 ▸ Chain Rule of Entropy

Let X_1^n be a tuple of any n discrete random variables supported on finite alphabets, then

$$H(X_1^n) = \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right).$$

Proof. The case where $n = 2$ follows directly from the result that

$$H(X_1,X_2) = H(X_1) + H(X_2 \mid X_1).$$

Suppose that there exists some integer $k \geq 2$ such that $H\left(X_1^k\right) = \sum_{i=1}^{k} H\left(X_i \mid X_1^{i-1}\right)$, consider

$$\begin{aligned} H\left(X_1^{k+1}\right) &= H\left(X_1^k, X_{k+1}\right) \\ &= H\left(X_1^k\right) + H\left(X_{k+1} \mid X_1^k\right) \\ &= \sum_{i=1}^{k} H\left(X_i \mid X_1^{i-1}\right) + H\left(X_{k+1} \mid X_1^k\right) \\ &= \sum_{i=1}^{k+1} H\left(X_i \mid X_1^{i-1}\right). \end{aligned}$$

□

A direct application of Proposition 2.1.6 yields the following result:

Corollary 2.1.7 ▸ Chain Rule of Entropy for Conditional Joint Distributions

Let X, Y, Z be discrete random variables supported on finite alphabets, then

$$H(X,Y \mid Z) = H(X \mid Z) + H(Y \mid X,Z).$$

Proof. Let $X_1 \coloneqq X \mid Z$ and $X_2 \coloneqq Y \mid Z$, then

$$\begin{aligned} H(X, Y \mid Z) &= H(X_1, X_2) \\ &= H(X_1) + H(X_2 \mid X_1) \\ &= H(X \mid Z) + H\big((Y \mid Z) \mid (X \mid Z)\big) \\ &= H(X \mid Z) + H(Y \mid X, Z). \end{aligned}$$

□

Given different distributions for the same random variable, we may be interested to know how much the distributions differ from one another. In other words, we wish to measure how much one distribution is different from another in terms of uncertainty.

Definition 2.1.8 ▸ Relative Entropy

Let p and q be probability mass functions for some discrete random variable X supported over an alphabet $\mathcal{X}$. The **relative entropy**, or alternatively, **Kullback-Leibler (KL) divergence**, between p and q is defined as

$$D(p \parallel q) \coloneqq \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}.$$

The above definition essentially describes the "difference" between two distributions as their expected ratio because

$$\sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}\left[\log_2 \frac{p(X)}{q(X)}\right].$$

Remark. Using a similar argument to Definition 2.1.2, we set the following conventions:

1. $0 \log_2 \frac{0}{q} = 0$ for all $q \in \mathbb{R}$;
2. $p \log_2 \frac{p}{0} = +\infty$ for all $p \in \mathbb{R}$.

Relative entropy can be defined in a conditional context as well.

Definition 2.1.9 ▸ Conditional Relative Entropy

Let $p_{X,Y}$ and $q_{X,Y}$ be joint probability mass functions for some pair of discrete random variables (X, Y) supported over an alphabet $\mathcal{X} \times \mathcal{Y}$. The **conditional relative entropy**

between $p_{Y|X}$ and $q_{Y|X}$ averaged over X is

$$D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \coloneqq \sum_{x \in \mathcal{X}} p_X(x) D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right).$$

Notice that by applying Proposition 2.1.6, we have

$$H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

due to $H(X, Y)$ being symmetric. This implies that

$$H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

Informally speaking, the left-hand side of the above identity is the remaining uncertainty of X after knowing Y, while the right-hand side is that of Y after knowing X. Note that this quantity can be interpreted as "the uncertain part of X and Y which cannot be reduced by knowing one of them". In other words, this remaining uncertainty is shared by both X and Y. The following notion formalises this observation:

Definition 2.1.10 ▸ Mutual Information

Let (X, Y) be a pair of discrete random variables with joint probability mass function $p_{X,Y}$. The **mutual information** between X and Y is defined as

$$I(X; Y) \coloneqq D\left(p_{X,Y} \parallel p_X \cdot p_Y\right).$$

It turns out that mutual information is symmetric, because

$$\begin{aligned} I(X; Y) &= D\left(p_{X,Y} \parallel p_X \cdot p_Y\right) \\ &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} \\ &= \mathbb{E}_{p_{X,Y}}\left[\log_2 \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)}\right]. \end{aligned}$$

One may check that

$$I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

In this way, by using Proposition 2.1.6, we can also see that

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

and hence the symmetric property of mutual information.

Naturally, the mutual information between X and Y cannot exceed the entropy of either of them. Therefore, it is intuitive that

$$0 \leq I(X;Y) \leq \min\left\{\log_2|\mathcal{X}|, \log_2|\mathcal{Y}|\right\}.$$

Mutual information can be conditional as well.

Proposition 2.1.11 ▸ Chain Rule of Mutual Information

Let (X_1^n, Y) be a tuple of discrete random variables with joint probability mass function p, then

$$I\left(X_1^n; Y\right) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_1^{i-1}\right).$$

Proof. By using Proposition 2.1.6,

$$\begin{aligned} I\left(X_1^n; Y\right) &= H\left(X_1^n\right) - H\left(X_1^n \mid Y\right) \\ &= \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right) - \sum_{i=1}^{n} H\left(X_i \mid Y, X_i^{i-1}\right) \\ &= \sum_{i=1}^{n} \left(H\left(X_i \mid X_1^{i-1}\right) - H\left(X_i \mid Y, X_i^{i-1}\right)\right) \\ &= \sum_{i=1}^{n} I\left(X_i; Y \mid X_1^{i-1}\right). \end{aligned}$$

□

We could make some analogy between entropy and set theory. Suppose we have two random variables X and Y, we could let some sets $\mathcal{H}_X$ and $\mathcal{H}_Y$ represent $H(X)$ and $H(Y)$. It is intuitive to see that $H(X \mid Y)$ corresponds to $\mathcal{H}_X \backslash \mathcal{H}_Y$, $H(X,Y)$ corresponds to $\mathcal{H}_X \cup \mathcal{H}_Y$, and that $I(X;Y)$ corresponds to $\mathcal{H}_X \cap \mathcal{H}_Y$.

This inspires us to study mutual information between more than 2 random variables via the principle of inclusion and exclusion. However, the situation becomes problematic when we consider more random variables. It can be shown that there exist random variables X, Y, Z such that $I(X;Y;Z) < 0$, which does not make much sense in information theory.

2.2 Information Inequality

A lot of theorems in information theory are developed from inequalities. Among them, the core inequality result is known as the *information inequality* which can be used to prove a

wide range of corollaries.

Theorem 2.2.1 ▶ Information Inequality

For any probability mass functions p and q for some random variable X, $D(p \parallel q) \geq 0$. The equality is attained if and only if $p = q$.

Proof. Let $A \coloneqq \{x \in \mathcal{X} : \ p(x) > 0\}$. Take $Y \coloneqq \frac{q(X)}{p(X)}$ with support

$$\mathcal{Y} \coloneqq \left\{ \frac{q(x)}{p(x)} : \ x \in A \right\}.$$

By Theorem 1.4.4, we have

$$\sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} = \mathbb{E}_p \left[\log_2 Y \right] \leq \log_2 \mathbb{E}_p [Y] = \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)}.$$

Note that $q(x) \geq 0$ for all $x \in \mathcal{X}$ and $p(x) > 0$ for all $x \in A$. Therefore,

$$\begin{aligned}
-D(p \parallel q) &= -\sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \\
&\leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\
&= \log_2 \sum_{x \in A} q(x) \\
&\leq \log_2 \sum_{x \in \mathcal{X}} q(x) \\
&= \log_2 1 \\
&= 0.
\end{aligned}$$

Therefore, $D(p \parallel q) \geq 0$. Clearly, the equality holds if and only if

$$\mathbb{E}_p \left[\log_2 Y \right] = \log_2 \mathbb{E}_p [Y] \quad \text{and} \quad \sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x).$$

Note that $f(x) = \log_2 x$ is strictly convex, so $\mathbb{E}_p \left[\log_2 Y \right] = \log_2 \mathbb{E}_p [Y]$ if and only if Y is constant, i.e., $\frac{q(x)}{p(x)} = c$ for some fixed $c \in \mathbb{R}$ for all $x \in A$. Notice that this is equivalent to

$$1 = \sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c,$$

i.e., $p(x) = q(x)$ for all $x \in A$. Note that in the same time, $q(x) = 0$ for all $x \in \mathcal{X} \backslash A$,

i.e., $q(x) = 0$ if and only if $p(x) = 0$. Therefore, $p = q$ as desired. □

The information inequality leads to many bounding conditions to the common quantities we have discussed so far.

Corollary 2.2.2 ▸ Mutual Information Is Non-negative

For any jointly distributed discrete random variables X and Y, $I(X;Y) \geq 0$ with equality attained if and only if X and Y are independent.

Proof. Notice that

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)\, p_Y(y) = \left(\sum_{x\in\mathcal{X}} p_X(x)\right)\left(\sum_{y\in\mathcal{Y}} p_Y(y)\right) = 1,$$

so $p_X \cdot p_Y$ is a probability mass function for (X, Y). Therefore, by Theorem 2.2.1,

$$I(X;Y) = D\left(p_{X,Y} \parallel p_X \cdot p_Y\right) \geq 0,$$

where the equality is attained if and only if $p_{X,Y} = p_X \cdot p_Y$, i.e., X and Y are independent. □

Naturally, the conditional relative entropy should also be non-negative.

Corollary 2.2.3 ▸ Conditional Relative Entropy Is Non-negative

For any pair of discrete random variables (X, Y), $D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \geq 0$ with equality attained if and only if $p_{Y|X}(\cdot \mid x) = q_{Y|X}(\cdot \mid x)$ for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$.

Proof. By Theorem 2.2.1,

$$D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right) \geq 0$$

for all $x \in \mathcal{X}$. Since $p_X(x) \geq 0$ for all $x \in \mathcal{X}$, clearly

$$D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \geq 0,$$

where the equality is attained if and only if

$$D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right) = 0$$

for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$. This is equivalent to $p_{Y|X}(\cdot \parallel x) = q_{Y|X}(\cdot \parallel x)$ for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$. □

We can do a similar argument for conditional mutual information as well.

Corollary 2.2.4 ▸ Conditional Mutual Information Is Non-negative

For any discrete random variables X, Y, Z, we have $I(X;Y \mid Z) \geq 0$ with equality attained if and only if X-Z-Y is a Markov chain.

Proof. Notice that by Theorem 2.2.1,

$$I(X;Y \mid Z) = D\left(p_{X,Y|Z} \parallel p_{X|Z} \cdot p_{Y|Z}\right) \geq 0,$$

where the equality is attained if and only if $p_{X,Y|Z} = p_{X|Z} \cdot p_{Y|Z}$. □

Recall that we previously mentioned that $0 \leq H(X) \leq \log_2 |\mathcal{X}|$. The upper bound can be derived using the information inequality as well.

Proposition 2.2.5 ▸ Upper Bound of Entropy

For any discrete random variable X supported on a finite alphabet $\mathcal{X}$, we have $H(X) \leq \log_2 |\mathcal{X}|$ with equality attained if and only if p_X is uniform on $\mathcal{X}$.

Proof. Define $u(x) \coloneqq \frac{1}{|\mathcal{X}|}$ to be the uniform distribution over $\mathcal{X}$. Consider

$$\begin{aligned} D(p_X \parallel u) &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{u(x)} \\ &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 |\mathcal{X}| p_X(x) \\ &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 |\mathcal{X}| + \sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) \\ &= \log_2 |X| - H(X). \end{aligned}$$

By Theorem 2.2.1, we have $D(p_X \parallel u) \geq 0$ and so $H(X) \leq \log_2 |X|$. The equality is attained if and only if $p_X = u$ is uniform over $\mathcal{X}$. □

One important result derived from this upper bound is as follows:

Corollary 2.2.6 ▸ Conditioning Does Not Increase Entropy

For any pair of discrete random variables (X, Y), we have $H(X \mid Y) \leq H(X)$ with equality attained if and only if X and Y are independent.

Proof. Notice that

$$H(X) - H(X \mid Y) = I(X;Y) \geq 0$$

by Corollary 2.2.2, so $H(X \mid Y) \leq H(X)$ as desired. □

However, do note that there could exist some $y \in \mathcal{Y}$ such that $H(X \mid Y = y) > H(X)$. For example, consider the following joint distribution:

Y \ X	1	2
1	0	0.75
2	0.125	0.125

We compute the conditional entropy values:

$$H(X \mid Y = 1) = -p_{X|Y}(2 \mid 1) \log_2 p_{X|Y}(2 \mid 1) = 0,$$
$$H(X \mid Y = 2) = -p_{X|Y}(1 \mid 2) \log_2 p_{X|Y}(1 \mid 2) - p_{X|Y}(2 \mid 2) \log_2 p_{X|Y}(2 \mid 2) = 1.$$

However, $H(X \mid Y) = p_Y(2) H(X \mid Y = 2) = 0.25 < H(X \mid Y = 2)$.

An important result which can be proven using the above facts is the *Hans's inequality*, which states that

$$H(X_1, X_2, X_3) \leq \frac{1}{2}\big(H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3)\big).$$

This inequality can be easily proven by considering the fact that

$$H(X_1, X_2, X_3) = H(X_1 \mid X_2, X_3) + H(X_2, X_3).$$

However, the inequality can in fact be seen as a special case for a more powerful result known as *Shearer's inequality*.

Theorem 2.2.7 ▸ Shearer's Inequality

For any $n \in \mathbb{N}$ and any random subset $\mathcal{S} \subseteq [n] \setminus \{0\}$ such that $\Pr(i \in \mathcal{S}) \geq \mu$ for all $i \in [n]$,

$$\mathbb{E}[H(X_{\mathcal{S}})] \geq H(X_1^n).$$

Proof. For any $\mathcal{S} \subseteq [n] \setminus \{0\}$, we can write

$$\mathcal{S} = \{i_1, i_2, \cdots, i_k\}$$

for some $k \in [n]$ such that $i_p < i_q$ whenever $p < q$. Notice that

$$H(X_S) = \sum_{j=1}^{k} H\left(X_{i_j} \mid X_{i_1}, X_{i_2}, \cdots, X_{i_{j-1}}\right) \geq \sum_{j=1}^{k} H\left(X_{i_j} \mid X_1^{i_j-1}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[H(X_S)\right] &\geq \mathbb{E}\left[\sum_{i \in S} H\left(X_i \mid X_1^{i-1}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}\{i \in S\} H\left(X_i \mid X_1^{i-1}\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}\{i \in S\}\right] H\left(X_i \mid X_1^{i-1}\right) \\
&= \sum_{i=1}^{n} \Pr(i \in S) H\left(X_i \mid X_1^{i-1}\right) \\
&\geq \mu \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right) \\
&= \mu H\left(X_1^n\right).
\end{aligned}$$

□

A different way to state Shearer's inequality uses a combinatorial construction:

Theorem 2.2.8 ▸ Equivalent Statement for Shearer's Inequality

Let $X_1, X_2, \cdots, X_n$ be n random variables for some $n \in \mathbb{N}^+$. For every positive integer $m \leq n$, define $\mathcal{S}_m \subseteq \mathcal{P}([n] \setminus \{0\})$ to be a collection of subsets of $[n] \setminus \{0\}$ such that for each $i \in [n] \setminus \{0\}$, there exist at least m sets containing i in $\mathcal{S}_m$, then

$$H(X_1, X_2, \cdots, X_n) \leq \frac{1}{m} \sum_{S \in \mathcal{S}_m} H(X_S),$$

where $X_S = \{X_i : i \in S \subseteq [n] \setminus \{0\}\}$.

Using either version, we can prove Hans's inequality by simply taking $n = 3$ and $\mathcal{S}$ to be uniform over all subsets of $\{1, 2, 3\}$ with cardinality 2.

Corollary 2.2.9 ▸ Hans's Inequality

For any random variables X_1, X_2, X_3,

$$H(X_1, X_2, X_3) \leq \frac{1}{2}\big(H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3)\big).$$

Theorem 2.2.7 has many seemingly surprising applications. For example, we can apply it to prove some results in graph theory by probabilistic methods.

Proposition 2.2.10 ▸ The Number of Triangles in Simple Undirected Graphs

Let G be a simple undirected graph containing t triangles, then

$$t \leq \frac{1}{6}\big(2e(G)\big)^{\frac{3}{2}}.$$

Proof. Let $\mathcal{V}$ be the set of all ordered 3-tuples inducing some $C_3 \subseteq G$. Notice that for every $H \subseteq G$ such that $H \cong C_3$, there are 6 different permutations of $V(H)$. Therefore, $|\mathcal{V}| = 6t$. Let (X_1, X_2, X_3) be uniformly distributed over $\mathcal{V}$, then clearly

$$\Pr\big(X_1 = v_i, X_2 = v_j, X_3 = v_k\big) = \frac{1}{6t}$$

for all $(v_i, v_j, v_k) \in \mathcal{V}$. By Proposition 2.2.5, it is clear that $H(X_1, X_2, X_3) = \log_2(6t)$. Let $\mathcal{S}$ be uniformly distributed over $\mathcal{P}(\{1, 2, 3\})$, then by Theorem 2.2.7,

$$\mathbb{E}\left[H(X_{\mathcal{S}})\right] \geq \frac{2}{3}H(X_1, X_2, X_3) = \frac{2}{3}\log_2(6t).$$

Notice that this implies that there exists some $\mathcal{T} \subseteq \{1, 2, 3\}$ with $|\mathcal{T}| = 2$ such that $H(X_{\mathcal{T}}) \geq \frac{2}{3}\log_2(6t)$. However, observe that $X_{\mathcal{T}}$ is supported over the set of all ordered pairs induced by $E(G)$, which means that $H(X_{\mathcal{T}}) \leq \log_2\big(2e(G)\big)$. Therefore,

$$\frac{2}{3}\log_2(6t) \leq \log_2\big(2e(G)\big).$$

One may check that this inequality is equivalent to

$$t \leq \frac{1}{6}\big(2e(G)\big)^{\frac{3}{2}}.$$

□

We have seen how useful Corollary 2.2.6 is. Actually, the result can be generalised further. We first introduce a preliminary definition.

Definition 2.2.11 ▸ Mutual Independence

Let $X_1, X_2, \cdots, X_n$ be any n random variables. They are said to be **mutually independent** if for any $S \subseteq [n] \setminus \{0\}$,

$$p_{X_S} = \prod_{i \in S} p_{X_i}.$$

Now, we propose the following inequality for conditional entropy:

Corollary 2.2.12 ▸ Sub-additivity of Conditional Entropy

For any random variables $X_1, X_2, \cdots, X_n$,

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

where the equality attained if and only if the X_i's are mutually independent.

Proof. By Corollary 2.2.6, $H\left(X_i \mid X_1^{i-1}\right) \leq H(X_i)$ for all $i = 1, 2, \cdots, n$, so

$$H(X_1^n) = \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right) \leq \sum_{i=1}^{n} H(X_i).$$

The equality is attained if and only if X_i and X_j are independent whenever $i \neq j$, i.e., the X_i's are mutually independent. □

The next inequality is very useful tool which can be used to prove the many results derived from the information inequality so far.

Theorem 2.2.13 ▸ Log-sum Inequality

Let $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ be non-negative real-valued sequences, then

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Proof. If there exists some $i \in \mathbb{N}^+$ such that $b_i = 0$, then the left-hand side is $+\infty$. If there exists some $i \in \mathbb{N}^+$ such that $a_i = 0$, then a_i contribute 0 to both side of the inequality. Therefore, without loss of generality, we can assume that $a_i, b_i > 0$ for all $i = 1, 2, \cdots, n$. Let $a = \sum_{i=1}^{n} a_i$ and $b = \sum_{i=1}^{n} b_i$. One may check that the

function $f(x) = x \log_2 x$ is strictly convex, so

$$\sum_{i=1}^{n} \frac{b_i a_i}{b b_i} \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} \frac{b_i a_i}{b b_i} \right) \log_2 \sum_{i=1}^{n} \frac{b_i a_i}{b b_i}.$$

Simplifying the inequality yields

$$\frac{1}{b} \sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \frac{a}{b} \log_2 \frac{a}{b},$$

and so

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq a \log_2 \frac{a}{b}.$$

□

One result which can be proven with the aid of the log-sum inequality is the joint convexity of relative entropy.

Proposition 2.2.14 ▸ Convexity of Relative Entropy

$D(p \parallel q)$ is jointly convex.

Proof. Let p_1, p_2, q_1, q_2 be probability mass functions for the same random variable X. It suffices to prove that for any $\lambda \in [0, 1]$,

$$D\big(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2\big) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2).$$

For any $x \in \mathcal{X}$, let

$$\begin{aligned} x_{p_1} &= \lambda p_1(x), & x_{p_2} &= (1-\lambda) p_2(x); \\ x_{q_1} &= \lambda q_1(x), & x_{q_2} &= (1-\lambda) q_2(x). \end{aligned}$$

By Theorem 2.2.13,

$$\big(x_{p_1} + x_{p_2}\big) \log_2 \frac{x_{p_1} + x_{p_2}}{x_{q_1} + x_{q_2}} \leq x_{p_1} \log_2 \frac{x_{p_1}}{x_{q_1}} + x_{p_2} \log_2 \frac{x_{p_2}}{x_{q_2}}.$$

□

We can use a similar approach to analyse the convexity of entropy

Proposition 2.2.15 ▸ Concavity of Entropy

For any discrete random variable X with a finite alphabet and distribution p, $H(p)$ is concave in p.

Proof. Let u be the uniform distribution for X, then $u(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$. Notice that

$$\begin{aligned} D(p \parallel u) &= \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{u(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} \\ &= \log_2 |\mathcal{X}| - H(p). \end{aligned}$$

By Proposition 2.2.14, $D(p \parallel u)$ is convex, so $H(p)$ must be concave. □

Alternatively, let $T \sim \text{Bernoulli}(\lambda)$ and define $Z \coloneqq X \mid T$, where $X \mid T = 1$ and $X \mid T = 0$ have distributions p_1 and p_2 respectively, then clearly $p_Z = \lambda p_1 + (1-\lambda) p_2$. Therefore,

$$\begin{aligned} H\big(\lambda p_1 + (1-\lambda) p_2\big) &= H(Z) \\ &\geq H(Z \mid T) \\ &= p_T(1) H(Z \mid T = 1) + p_T(0) H(Z \mid T = 0) \\ &= \lambda H(X \mid T = 1) + (1-\lambda) H(X \mid T = 0) \\ &= \lambda H(p_1) + (1-\lambda) H(p_2). \end{aligned}$$

This gives a more classical approach to proving Proposition 2.2.15.

Note that we can view $I(X;Y)$ as a function of $p_{X,Y}$, which can be further unpacked as a function of p_X and $p_{Y|X}$. Here, p_X is known as the *input distribution* and $p_{Y|X}$ is known as the *channel*.

Theorem 2.2.16 ▸ Convexity of Mutual Information

$I(X;Y)$ is concave in p_X and convex in $p_{Y|X}$.

Proof. Fix $p_{Y|X}$, consider

$$\begin{aligned} I(X;Y) &= H(Y) - H(X \mid Y) \\ &= H(p_Y) - \sum_{x \in \mathcal{X}} p_X(x) H(Y \mid X = x). \end{aligned}$$

Note that $H(Y \mid X = x)$ only depends on $p_{Y|X}$ and so is a constant and for any $y \in \mathcal{Y}$,

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_X(x)\, p_{Y|X}(y \mid x)$$

is linear in $p_X(x)$. Therefore, $I(X;Y)$ is linear in p_X and so concave in p_X. Now, fix p_X and define $T \sim \text{Bernoulli}(\lambda)$ to be independent of X, then

$$p_{Y|X} = \lambda p_{Y|X,T=1} + (1-\lambda)\, p_{Y|X,T=0}.$$

Notice that

$$I(X;T) + I(X;Y \mid T) = I(X;Y) = I(X;Y) + I(X;T \mid Y).$$

Since $I(X;T) = 0$, this implies that $I(X;Y \mid T) = I(X;Y) + I(X;T \mid Y)$. This means that

$$\begin{aligned} I(X;Y) &\leq I(X;Y \mid T) \\ &= \lambda I(X;Y \mid T=1) + (1-\lambda)\, I(X;Y \mid T=0), \end{aligned}$$

which implies that $I(X;Y)$ is convex in $p_{Y|X}$. □

The next inequality is concerning data processing.

Theorem 2.2.17 ▸ Data Processing Inequality (DPI)

If X-Y-Z forms a Markov chain, then $I(X;Y) \geq I(X;Z)$.

Proof. Notice that

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

Since X-Y-Z forms a Markov chain, $I(X;Z \mid Y) = 0$. Therefore,

$$I(X;Y) = I(X;Z) + I(X;Y \mid Z) \geq I(X;Z).$$

□

Mathematically, data processing can be described as mapping a random variable Y to some transformed image via a function g. Recall that in Proposition 1.2.4, we have shown that for any random variables X and Y, X-Y-$f(Y)$ is always a Markov chain, which motivates the following application of the DPI.

Corollary 2.2.18 ▸ Mutual Information Does Not Increase After Processing

For any random variables X and Y, we have $I(X;Y) \geq I\big(X; g(Y)\big)$ for all function g.

One implication of this is that no matter what method is used to process the information Y, the shared knowledge between Y and some other data set X can be at best retained at the same level as before.

2.3 Sufficient Statistics

Consider a parametric family of probability distributions $\{f_\theta(x) : \theta \in \Theta\}$ for some index set Θ. Let $T(X)$ be any statistic. One may check that Θ-X-$T(X)$ is a Markov chain. By Theorem 2.2.17 we know that $I(\Theta; T(X)) \leq I(\Theta; X)$. If the equality is attained, no information is lost in this statistic.

Definition 2.3.1 ▸ Sufficient Statistic

A function $T : \mathcal{X} \to \mathbb{R}$ is said to be a **sufficient statistic** relative to a parametric family $\mathcal{F}_\theta := \{f_\theta(x) : \theta \in \Theta\}$ of probability distributions if Θ-$T(X)$-X forms a Markov chain.

Let X be any random variable and Y be another random variable correlated to Y. Suppose now we wish to estimate X via observations about Y. Let $\widehat{X}(Y)$ be an estimator obtained this way about X.

If $H(X \mid Y) = 0$, one can expect that a perfect estimation is possible. On the other hand, if $H(X \mid Y) = \log_2 |\mathcal{X}|$, the the estimation is bad. Notice that this happens if and only if X is uniform and independent of Y. In reality, we may wish $H(X \mid Y)$ to be small, where the error of estimation can be small.

Theorem 2.3.2 ▸ Fano's Inequality

For any estimator $\widehat{X}$ obtained from Y, let $p_e := \Pr\big(\widehat{X} \neq X\big)$ be the probability of error, then

$$H_b(p_e) + p_e \log_2 |\mathcal{X}| \geq H\big(X \mid \widehat{X}\big) \geq H(X \mid Y).$$

Proof. Define $E := \mathbf{1}\big\{\widehat{X} \neq X\big\}$ to be the error random variable, then $p_e = \Pr(E = 1)$. Consider

$$H\big(E \mid \widehat{X}\big) + H\big(X \mid E, \widehat{X}\big) = H\big(E, X \mid \widehat{X}\big) = H\big(X \mid \widehat{X}\big) + H\big(E \mid X, \widehat{X}\big).$$

It is clear that $H\left(E \mid X, \widehat{X}\right) = 0$. By Corollary 2.2.6, since E is a Bernoulli random variable, we have

$$H\left(E \mid \widehat{X}\right) \leq H(E) = H_b(p_e).$$

Note that

$$H\left(X \mid E, \widehat{X}\right) = \Pr(E=1) H\left(X \mid E=1, \widehat{X}\right) + \Pr(E=0) H\left(X \mid E=0, \widehat{X}\right).$$

Clearly, $H\left(X \mid E=0, \widehat{X}\right) = 0$ and $H\left(X \mid E=1, \widehat{X}\right) \leq \log_2 |\mathcal{X}|$, so

$$H\left(X \mid E, \widehat{X}\right) \leq p_e \log_2 |\mathcal{X}|.$$

Therefore,

$$H\left(X \mid \widehat{X}\right) = H\left(E \mid \widehat{X}\right) + H\left(X \mid E, \widehat{X}\right) \leq H_b(p_e) + p_e \log_2 |\mathcal{X}|.$$

Note that $\widehat{X}$ is a function of Y, so by Corollary 1.2.4, X-Y-$\widehat{X}$ forms a Markov chain. By Theorem 2.2.17,

$$\begin{aligned} H(X) - H(X \mid Y) &= I(X;Y) \\ &\geq I\left(X; \widehat{X}\right) \\ &= H(X) - H\left(X \mid \widehat{X}\right), \end{aligned}$$

which reduces to $H\left(X \mid \widehat{X}\right) \geq H(X \mid Y)$. □

With some algebraic manipulations, it can be obtained from Theorem 2.3.2 that

$$p_e \geq \frac{H(X \mid Y) - 1}{\log_2 |\mathcal{X}|}.$$

In other words, this means that

$$\inf \Pr\left(\widehat{X}(Y) \neq X\right) \geq \frac{H(X \mid Y) - 1}{\log_2 |\mathcal{X}|}.$$

This is one of the few results which offer a **lower bound estimate** for probabilities. In particular, this result shows that in a non-trivial scenario ($|\mathcal{X}| > 1$), perfect estimator is attainable only when $H(X \mid Y) \leq 1$.

3

Data Compression

3.1 Asymptotic Equipartition Property

In this section, we try to investigate some bounding conditions on the rate of data compression. We focus on a special type of sources called the *discrete memoryless source* (DMS), which can be viewed as a sequence of **independent and identically distributed** discrete random variables such that the distribution of X_n is independent of the distributions of all X_i's with $i < n$.

Definition 3.1.1 ▸ Discrete Memoryless Source

A **discrete memoryless source** is a sequence of identically distributed random variables X_1^n such that X_i is independent of all X_j whenever $j < i$.

One may check that a DMS is mutually independent.

We introduce the following notion:

Definition 3.1.2 ▸ ϵ-Weakly Typical Set

Let X be a discrete memoryless random variable with distribution p_X. An **ϵ-Weakly Typical Set** of X is defined as

$$A_\epsilon^{(n)}(X) \coloneqq \left\{ \boldsymbol{x} \in \mathcal{X}^n : \left| \frac{1}{n} \log_2 \frac{1}{\prod_{i=1}^n p_X(x_i)} - H(X) \right| \leq \epsilon \right\}$$

where $n \in \mathbb{N}^+$.

While this definition could seem obscure, we can unpack the inequality into

$$2^{-n\left(H(X)+\epsilon\right)} \leq \prod_{i=1}^n p_X(x_i) \leq 2^{-n\left(H(X)-\epsilon\right)}.$$

As $\epsilon \to 0$, we see that $A_\epsilon^{(n)}(X)$ defines a set of sequences which have an asymptotically equal probability of occurring as the output from a discrete memoryless source.

Theorem 3.1.3 ▸ Asymptotic Equipartition Property

If X_1^n is a discrete memoryless source, then for all $\epsilon > 0$, there exists some $N \in \mathbb{N}^+$ such that for all $n > N$,

$$\Pr\left(X_1^n \in A_\epsilon^{(n)}(X)\right) \geq 1 - \epsilon,$$

where

$$(1-\epsilon)\, 2^{n\left(H(X)-\epsilon\right)} \leq \left|A_\epsilon^{(n)}(X)\right| \leq 2^{n\left(H(X)+\epsilon\right)}.$$

Proof. Consider

$$\begin{aligned}
\Pr\left(X_1^n \notin A_\epsilon^{(n)}(X)\right) &= \Pr\left(\left|\frac{1}{n}\log_2 \frac{1}{\prod_{i=1}^n p_X(Z_i)} - H(X)\right| > \epsilon\right) \\
&= \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{1}{p_X(X_i)} - H(X)\right| > \epsilon\right) \\
&= \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{1}{p_X(X_i)} - \mathbb{E}\left[\frac{1}{p_X(X_1)}\right]\right| > \epsilon\right).
\end{aligned}$$

Since the X_i's are identical and independent, by Theorem 1.3.3, for all $\epsilon > 0$, there exists some $N \in \mathbb{N}^+$ such that for all $n > N$, we have

$$\Pr\left(X_1^n \notin A_\epsilon^{(n)}(X)\right) < \epsilon,$$

which implies that

$$\Pr\left(X_1^n \in A_\epsilon^{(n)}(X)\right) = 1 - \Pr\left(X_1^n \notin A_\epsilon^{(n)}(X)\right) \geq 1 - \epsilon.$$

Notice that for each $\boldsymbol{x} \in A_\epsilon^{(n)}(X)$, we have

$$2^{-n\left(H(X)+\epsilon\right)} \leq \prod_{i=1}^n p_X(x_i) \leq 2^{-n\left(H(X)-\epsilon\right)}.$$

Since $A_\epsilon^{(n)}(X) \subseteq \mathcal{X}^n$, we have

$$\begin{aligned}
1 &\geq \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}(X)} \prod_{i=1}^n p_X(x_i) \\
&\geq \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}(X)} 2^{-n\left(H(X)+\epsilon\right)} \\
&= 2^{-n\left(H(X)+\epsilon\right)} \left|A_\epsilon^{(n)}(X)\right|.
\end{aligned}$$

Therefore,

$$\left|A_\epsilon^{(n)}(X)\right| \leq 2^{n\left(H(X)+\epsilon\right)}.$$

For all $n > N$, consider

$$\begin{aligned}
1-\epsilon &\leq \Pr\left(X_1^n \in A_\epsilon^{(n)}(X)\right) \\
&= \sum_{\boldsymbol{x}\in A_\epsilon^{(n)}(X)} \prod_{i=1}^{n} p_X(x_i) \\
&\leq \sum_{\boldsymbol{x}\in A_\epsilon^{(n)}(X)} 2^{-n\left(H(X)-\epsilon\right)} \\
&= 2^{-n\left(H(X)-\epsilon\right)}\left|A_\epsilon^{(n)}(X)\right|.
\end{aligned}$$

Therefore,

$$(1-\epsilon)\,2^{n\left(H(X)-\epsilon\right)} \leq \left|A_\epsilon^{(n)}(X)\right|.$$

□

Remark. The upper bound for $\left|A_\epsilon^{(n)}(X)\right|$ holds for all $n \in \mathbb{N}^+$, where the lower bound only holds when n is sufficiently large.

One important implication of the AEP is as follows: in computers, we code information into binary strings. Now, suppose X is a binary random variable, then clearly $\mathcal{X}^n = \{0,1\}^n$. By Theorem 3.1.3, we know that when an input binary string X_1^n is very long, there exists a very small subset $A_\epsilon^{(n)}(X) \subseteq \mathcal{X}^n$ of size approximately $2^{nH(X)}$. This is because if X is not uniform, then $0 \leq H(X) < 1$ and so $\left|A_\epsilon^{(n)}(X)\right|$ is much smaller than $|\mathcal{X}^n| = 2^n$. However, the probability that X_1^n falls in $A_\epsilon^{(n)}(X)$ is very high according to Theorem 3.1.3.

3.2 Fixed-to-Fixed-Length Data Compression

Consider a sequence of random variables X_1^n, we wish to construct an *encoder*

$$\text{Enc}: \ \mathcal{X}^n \to \{0,1\}^{nR}$$

for some $R \in \mathbb{R}$ to map any possible input to a binary string. Moreover, we also wish to construct a *decoder*

$$\text{Dec}: \ \{0,1\}^{nR} \to \widehat{\mathcal{X}}_1^n$$

to restore the original input as a sequence of estimators. This process is known as *compression*. This can be formalised as the following:

Definition 3.2.1 ▸ Fixed-To-Fixed-Length Source Code

An $\left(n, 2^{nR}\right)$-**fixed-to-fixed-length source code** for a random variable X is an **encoder**

$$f: \mathcal{X}^n \to \left\{1, 2, \cdots, 2^{nR}\right\}$$

plus a **decoder**

$$\varphi: \left\{1, 2, \cdots, 2^{nR}\right\} \to \mathcal{X}^n,$$

where n is known as the **block length** and R, **code rate**. We define $M \coloneqq f\left(X_1^n\right)$ as the **compression index** and $\widehat{X}_1^n \coloneqq \varphi(M)$ as the **reconstructed source**.

Remark. Technically, nR might not be an integer, but for the sake of simplicity, we take 2^{nR} to be the same as $\lceil 2^{nR} \rceil$.

Ideally, we wish to achieve perfect compression with $\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 0$, but that means that we need $R = \log_2 |X|$, which is as if we are not compressing any data contained in X. Therefore, we can relax this constraint a bit and now aim at

$$\lim_{n\to\infty} \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 0.$$

Definition 3.2.2 ▸ Achievable Rate

Let X be a random variable. A code rate $R \geq 0$ is said to be **achievable** if there exists a sequence of $\left(n, 2^{nR}\right)$-codes such that

$$\lim_{n\to\infty} \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 0.$$

In other words, as $n \to \infty$, we have $\widehat{X}_1^n = \varphi\left(f\left(X_1^n\right)\right) \approx X_1^n$. This is known as *vanishing probability of error*. The question is: how large should we take R to be in order to achieve this?

First, observe that if R is achievable, then any $R' > R$ is also achievable. Therefore, it suffices to focus on the minimal achievable rate.

Definition 3.2.3 ▸ Optimal Source Coding Rate

The **optimal source coding rate** for a discrete memoryless source X is defined as

$$R^*(X) \coloneqq \inf\{R \geq 0: R \text{ is achievable}\}.$$

The following theorem states the main result for fixed-to-fixed-length data compression.

Theorem 3.2.4 ▸ Fixed-To-Fixed-Length Data Compression

For any discrete memoryless source X, we have $R^*(X) = H(X)$.

Proof. By Theorem 3.1.3, $A_\epsilon^{(n)}(X)$ is finite for all $n \in \mathbb{N}^+$ and all $\epsilon > 0$, so we can always fix a bijection $m_n : A_\epsilon^{(n)}(X) \to \left\{1, 2, \cdots, \left|A_\epsilon^{(n)}(X)\right|\right\}$. For any $n \in \mathbb{N}^+$ and any $\epsilon > 0$, define an encoder $f_n : \mathcal{X}^n \to \left\{1, 2, \cdots, 2^{nR}\right\}$ by

$$f_n(X_1^n) = \begin{cases} m_n(X_1^n) & \text{if } X_1^n \in A_\epsilon^{(n)}(X) \\ 1 & \text{otherwise} \end{cases}$$

and a decoder $\varphi_n : \left\{1, 2, \cdots, 2^{nR}\right\} \to \mathcal{X}^n$ by $\varphi_n(M) = m_n^{-1}(M)$. Notice that by Theorem 3.1.3, $\left|A_\epsilon^{(n)}(X)\right| \leq 2^{n\left(H(X)+\epsilon\right)}$, so

$$R \leq H(X) + \epsilon$$

for all $\epsilon > 0$. Observe that for any X_1^n, there could be an error if $X_1^n \notin A_\epsilon^{(n)}(X)$, so

$$\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \leq \Pr\left(X_1^n \notin A_\epsilon^{(n)}(X)\right).$$

By Theorem 3.1.3, $\lim_{n\to\infty} \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 0$ as desired. This means that R is achievable and so $R^*(X) \leq H(X) + \epsilon$ for all $\epsilon > 0$.

Let $M = f_n(X_1^n)$. Since $\widehat{\mathcal{X}}_1^n$ is an estimator for $\mathcal{X}_1^n$ obtained from M, by Theorem 2.3.2 we have

$$\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \geq \frac{H(X_1^n \mid M) - 1}{\log_2\left(|\mathcal{X}|^n\right)}.$$

Therefore,

$$H(X_1^n \mid M) \leq n \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \log_2 |\mathcal{X}| + 1.$$

Notice that

$$\begin{aligned} nR &= \log_2 |M| \\ &\geq H(M) \\ &= H(M) - H(M \mid X_1^n) + H(M \mid X_1^n) \\ &= I(M; X_1^n) \\ &= nH(X) - H(X_1^n \mid M). \end{aligned}$$

By Corollary 2.2.12, $H(X_1^n) \leq nH(X)$, so we have

$$nR \geq nH(X) - n\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right)\log_2|\mathcal{X}| - 1.$$

Therefore,

$$R \geq H(X) - \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right)\log_2|\mathcal{X}| - \frac{1}{n}.$$

Since

$$\lim_{n\to\infty}\left(\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right)\log_2|\mathcal{X}| + \frac{1}{n}\right) = 0,$$

we have $R \geq H(X)$. Therefore, $R^*(X) \geq H(X)$. Combining the two directions we have $R^*(X) = H(X)$. □

We can attempt to relax the condition on the error probability to

$$\limsup_{n\to\infty}\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \leq \epsilon$$

for some $\epsilon \in [0,1)$ instead. This kind of behaviour is called ϵ-achievability. If we define the optimal ϵ-achievable rate to be

$$R_\epsilon^*(X) \coloneqq \{R \geq 0 : \ R \text{ is } \epsilon\text{-achievable}\}.$$

It is clear that $R^*(X) = R_0^*(X) \geq R_\epsilon^*(X)$ for all $\epsilon \in [0,1)$. Surprisingly, this does not improve $R^*(X)$. We first consider a lemma which helps us prove this:

Lemma 3.2.5 ▸ Han-Verdú Technique

For any $(n, 2^{nR})$-code,

$$\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \geq \sup_{\gamma>0}\left\{\Pr\left(\frac{1}{n}\log_2\frac{1}{\prod_{i=1}^n p_X(x_i)} \geq R+\gamma\right) - \mathrm{e}^{-n\gamma}\right\}.$$

Proof. For every $\gamma > 0$ and $n \in \mathbb{N}^+$, define

$$T_n \coloneqq \left\{\boldsymbol{x} \in \mathcal{X}^n : \ \frac{1}{n}\log_2\frac{1}{\prod_{i=1}^n p_X(x_i)} \geq R+\gamma\right\},$$
$$S_n \coloneqq \left\{\boldsymbol{x} \in \mathcal{X}^n : \ \varphi_n\big(f_n(\boldsymbol{x})\big) = \boldsymbol{x}\right\}.$$

Consider

$$\begin{aligned}\Pr\left(X_1^n \in T_n\right) &= \Pr\left(X_1^n \in T_n \cap S_n\right) + \Pr\left(X_1^n \in T_n \cap S_n^c\right) \\ &\leq \Pr\left(X_1^n \in T_n \cap S_n\right) + \Pr\left(X_1^n \in S_n^c\right) \\ &= \Pr\left(X_1^n \in T_n \cap S_n\right) + \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right).\end{aligned}$$

For each $\boldsymbol{x} \in T_n$, we have

$$\prod_{i=1}^{n} p_X\left(x_i\right) \leq 2^{-n(R+\gamma)},$$

and so

$$\begin{aligned}\Pr\left(X_1^n \in T_n \cap S_n\right) &= \sum_{\boldsymbol{x} \in T_n \cap S_n} \prod_{i=1}^{n} p_X\left(x_i\right) \\ &\leq \sum_{\boldsymbol{x} \in T_n \cap S_n} 2^{-n(R+\gamma)} \\ &\leq 2^{-n(R+\gamma)} |S_n| \\ &= 2^{-n\gamma}.\end{aligned}$$

Therefore, $\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \geq \Pr\left(X_1^n \in T_n\right) - 2^{-n\gamma} \geq \Pr\left(X_1^n \in T_n\right) - \mathrm{e}^{-n\gamma}$, and so

$$\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) \geq \sup_{\gamma > 0} \left\{ \Pr\left(\frac{1}{n} \log_2 \frac{1}{\prod_{i=1}^{n} p_X\left(x_i\right)} \geq R + \gamma \right) - \mathrm{e}^{-n\gamma} \right\}.$$

□

Using the above lemma, we can prove that an optimal achievable rate can never be smaller than the entropy of the source.

Theorem 3.2.6 ▸ Strong Converse of Optimal Rate

If $R < H\left(X\right)$, then $\lim_{n \to \infty} \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 1$.

Proof. For any $\eta > 0$, let $R = H\left(X\right) - \eta$, then by Lemma 3.2.5,

$$\begin{aligned}\Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) &\geq \Pr\left(\frac{1}{n} \log_2 \frac{1}{\prod_{i=1}^{n} p_X\left(x_i\right)} \geq R + \frac{\eta}{2} \right) - \mathrm{e}^{-n\frac{\eta}{2}} \\ &= \Pr\left(\frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X\left(x_i\right)} \geq H\left(X\right) - \frac{\eta}{2} \right) - \mathrm{e}^{-n\frac{\eta}{2}} \\ &\geq \Pr\left(\left| H\left(X\right) - \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X\left(x_i\right)} \right| \leq \frac{\eta}{2} \right) - \mathrm{e}^{-n\frac{\eta}{2}}.\end{aligned}$$

By Theorem 1.3.3,

$$\lim_{n\to\infty} \Pr\left(\left|H(X) - \frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)}\right| \leq \frac{\eta}{2}\right) = 1.$$

Therefore,

$$\lim_{n\to\infty} \Pr\left(X_1^n \neq \widehat{\mathcal{X}}_1^n\right) = 1.$$

□

3.3 Stochastic Processes

In the previous section, we have discussed some results in fixed-to-fixed-length source coding for discrete memoryless sources. However, there are times where the source is not memoryless. A general model we adopt for such cases is *stochastic processes*.

Definition 3.3.1 ▸ Discrete Time Stochastic Process

A **discrete time stochastic process** is a sequence of random variables $\{X_i\}_{i\in\mathbb{N}^+}$ such that

$$\sum_{x_n\in\mathcal{X}_n} p_{X_1^n}(x_1, x_2, \cdots, x_n) = p_{X_1^{n-1}}(x_1, x_2, \cdots, x_{n-1})$$

for all $n \in \mathbb{N}^+$.

In a sense, a stochastic process is fully characterised by all of its joint PMFs such that the PMFs stay consistent under marginalisation. In our case, we focus on some stochastic processes satisfying certain special properties.

Definition 3.3.2 ▸ Stationary Stochastic Process

A stochastic process $X := \{X_i\}_{i\in\mathbb{N}^+}$ is **stationary** if for all $n \in \mathbb{N}^+$ and all $\ell \in \mathbb{N}$,

$$p_{X_1^n}(x_1, x_2, \cdots, x_n) = p_{X_{1+\ell}^{n+\ell}}(x_1, x_2, \cdots, x_n)$$

for all $x_1, x_2, \cdots, x_n \in \mathcal{X}$.

Here, ℓ is known as the *shift*. A stochastic can become a Markov chain if every random variable X_i in the sequence only depends on the immediate previous random variable.

Definition 3.3.3 ▸ Markov Chain

A stochastic process $X \coloneqq \{X_i\}_{i \in \mathbb{N}^+}$ is a **Markov chain** if

$$p_{X_{n+1}|X_1^n}(x_{n+1} \mid x_1, x_2, \cdots, x_n) = p_{X_{n+1}|X_n}(x_{n+1} \mid x_n)$$

for all $n \in \mathbb{N}^+$.

It can be easily verified that for such a Markov chain,

$$p_{X_1^n}(x_1, x_2, \cdots, x_n) = p_{X_1}(x_1) \prod_{i=1}^{n-1} p_{X_{i+1}|X_i}(x_{i+1} \mid x_i)$$

for all $n \in \mathbb{N}^+$. Now, the next question is: does the conditional probability

$$\Pr(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

depend on where we are in the chain? In general, the answer is "yes".

Definition 3.3.4 ▸ Time-Invariant Markov Chain

A markov chain $X \coloneqq \{X_i\}_{i \in \mathbb{N}^+}$ is **time-invariant** if

$$p_{X_{n+1}|X_n}(x_{n+1} \mid x_n) = p_{X_2|X_1}(x_2 \mid x_1)$$

for all $n \in \mathbb{N}^+$ and all $x_{n+1}, x_n \in \mathcal{X}$.

For time-invariant discrete-time markov chains, we can capture the transition conditional probabilities $\Pr(X_{n+1} = j \mid X_n = i)$ into a matrix.

Definition 3.3.5 ▸ Transition Probability Matrix

Let $X \coloneqq \{X_i\}_{i \in \mathbb{N}^+}$ be a time-invariant discrete-time markov chain. The **transition probability matrix** (TPM) $\boldsymbol{P}$ of X is defined such that

$$P_{ij} \coloneqq \Pr(X_{n+1} = j \mid X_n = i)$$

for all $n \in \mathbb{N}^+$ and all $i, j \in \mathcal{X}$.

Suppose that $\mathcal{X}$ is finite. Then clearly, with respect to the TPM, we can write

$$\begin{aligned}\Pr(X_{n+1} = j) &= \sum_{i\in\mathcal{X}} \Pr(X_{n+1} = j \mid X_n = i)\Pr(X_n = i)\\ &= \sum_{i\in\mathcal{X}} P_{ij}\Pr(X_n = i)\end{aligned}$$

for all $n \in \mathbb{N}^+$ and all $j \in \mathcal{X}$. Now, by writing everything in terms of matrix multiplication, we have

$$\begin{bmatrix} p_{X_{n+1}}(x_1) & p_{X_{n+1}}(x_2) & \cdots & p_{X_{n+1}}(x_{|\mathcal{X}|})\end{bmatrix} = \begin{bmatrix} p_{X_n}(x_1) & p_{X_n}(x_2) & \cdots & p_{X_n}(x_{|\mathcal{X}|})\end{bmatrix}\boldsymbol{P}.$$

Let us define

$$\pi_n \coloneqq \begin{bmatrix} p_{X_n}(x_1) & p_{X_n}(x_2) & \cdots & p_{X_n}(x_{|\mathcal{X}|})\end{bmatrix},$$

then π_n is clearly a distribution for X_n. Now, this equation seems suspicious enough that π_n might possess some sort of limiting behaviour, which we define as follows:

Definition 3.3.6 ▸ Stationary Distribution

Let $X \coloneqq \{X_i\}_{i\in\mathbb{N}^+}$ be a Markov chain with transition probability matrix $\boldsymbol{P}$. A distribution π on $\mathcal{X}$ is said to be a **stationary distribution** if $\pi = \pi\boldsymbol{P}$.

A fact is that not all Markov chains have a well-defined stationary distribution. The following two important properties for Markov chains help guarantee the existence of such a distribution.

1. *Irreducible*: a Markov chain is irreducible if it is possible to go from any state to any other state, i.e.,

$$\Pr(X_m = a \mid X_n = b) > 0, \qquad \Pr(X_m = b \mid X_n = a) > 0$$

   for all $a \neq b \in \mathcal{X}$ and $m > n$.

2. *Aperiodic*: a Markov chain is aperiodic if for any state $i \in \mathcal{X}$, either it is impossible to go back to i from i, or the greatest common divisor for the lengths of all paths going from i to itself is 1. Formally, this means that

$$\gcd\{n \in \mathbb{N}^+ : \Pr(X_n = i \mid X_0 = i) > 0\} = 1.$$

Theorem 3.3.7 ▸ Characterisation of Irreducible and Aperiodic Markov Chains

An irreducible and aperiodic Markov chain has a unique stationary distribution.

Suppose the stationary distribution π exists, then it is clear that we can find π by solving the system

$$\pi \boldsymbol{P} - \boldsymbol{I} = \boldsymbol{0}.$$

Of course, one might also attempt to diagonalise $\boldsymbol{P}$ to find $\lim_{n\to\infty} \boldsymbol{P}^n$ because $\pi_n = \pi_0 \boldsymbol{P}^n$. Furthermore, suppose

$$\pi = \left[p_1, p_2, \cdots, p_{|\mathcal{X}|} \right].$$

Notice that

$$p_j = \sum_{i \in \mathcal{X}} P_{ij} p_i$$

for all $j \in \mathcal{X}$. It is not difficult to see that $P_{ij} = p_j$ for all $i \in \mathcal{X}$, and so

$$\lim_{n\to\infty} \boldsymbol{P}^n = \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}.$$

Definition 3.3.8 ▸ Entropy Rate

Let $X \coloneqq \{X_i\}_{i \in \mathbb{N}^+}$ be a stochastic process. The **entropy rate** of X is defined as

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

if the limit exists.

There is an alternative way to define the entropy rate:

Definition 3.3.9 ▸ Prime Definition for Entropy Rate

Let $X \coloneqq \{X_i\}_{i \in \mathbb{N}^+}$ be a stochastic process. The **prime entropy rate** of X is defined as

$$H'(X) \coloneqq \lim_{n\to\infty} H\left(X_n \mid X_1^{n-1}\right)$$

if the limit exists.

In general, the two types of entropy rate may not be equivalent. However, we can check that they are the same for stationary processes.

Theorem 3.3.10 ▸ Equivalence of Entropy Rates for Stationary Sources

If $X \coloneqq \{X_i\}_{i\in\mathbb{N}^+}$ is a stationary stochastic process, then $H(X)$ and $H'(X)$ both exist and $H(X) = H'(X)$.

Proof. Since X is stationary,

$$\begin{aligned} H\left(X_n \mid X_1^{n-1}\right) &\leq H\left(X_n \mid X_2^{n-1}\right) \\ &= H\left(X_{n-1} \mid X_1^{n-2}\right). \end{aligned}$$

Note that $H\left(X_n \mid X_1^{n-1}\right) \geq 0$ for all $n \in \mathbb{N}^+$, so by monotone convergence theorem,

$$H'(X) = \lim_{n\to\infty} H\left(X_n \mid X_1^{n-1}\right)$$

exists. Notice that

$$\frac{1}{n} H(X_1, X_2, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H\left(X_n \mid X_1^{n-1}\right),$$

so clearly

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) = \lim_{n\to\infty} H\left(X_n \mid X_1^{n-1}\right) = H'(X).$$

□

We now state a theorem without proof which will generalise Theorem 3.1.3 to stationary stochastic processes.

Theorem 3.3.11 ▸ Shannon-McMillian-Breiman Theorem

For a stationary ergodic process X,

$$\lim_{n\to\infty} -\frac{1}{n} \log_2 p(X_1, X_2, \cdots, X_n) = H(X).$$

Now, we apply this theorem to Markov chains. We first state the following fact:

Proposition 3.3.12 ▸ Entropy Rate of Stationary Markov Chains

For any time-invariant stationary Markov chain X with transition probability matrix $\boldsymbol{P}$ and stationary distribution π, the entropy rate is given by

$$H(X) = H'(X) = H(X_2 \mid X_1)$$

where $X_1 \sim \pi$ and $X_2 \mid X_1$ follows the distribution induced by $\boldsymbol{P}$.

Proof. By Theorem 3.3.10, since X is a Markov chain,

$$\begin{aligned}
H(X) &= H'(X) \\
&= \lim_{n\to\infty} H\left(X_n \mid X_1^{n-1}\right) \\
&= \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\
&= \lim_{n\to\infty} \sum_{i\in\mathcal{X}} p_{X_{n-1}}(i) H(X_n \mid X_{n-1} = i) \\
&= \lim_{n\to\infty} \sum_{i\in\mathcal{X}} p_{X_{n-1}}(i) \left(- \sum_{j\in\mathcal{X}} p_{X_n|X_{n-1}}(j \mid i) \log_2 p_{X_n|X_{n-1}}(j \mid i) \right) \\
&= \lim_{n\to\infty} \sum_{i\in\mathcal{X}} p_{X_{n-1}}(i) \left(- \sum_{j\in\mathcal{X}} P_{ij} \log_2 P_{ij} \right) \\
&= \lim_{n\to\infty} \sum_{i\in\mathcal{X}} p_{X_{n-1}}(i) H(X_2 \mid X_1 = i) \\
&= \sum_{i\in\mathcal{X}} \pi(i) H(X_2 \mid X_1 = i) \\
&= H(X_2 \mid X_1).
\end{aligned}$$

□

Let us discuss a few applications of Markov chains.

3.3.1 Application: List Shuffling

Suppose we have a list containing k distinct items and we wish to shuffle the list such that the end outcome is **completely random**, i.e., the shuffled list is independent of the starting state of the original list. Let us make a naïve attempt first:

One-at-a-Time Shuffling

Let $X \coloneqq \{X_i\}_{i\in\mathbb{N}^+}$ be a stochastic process where X_i represents the permutation of the list after the i-th shuffle. First, observe that $|\mathcal{X}| = k!$. Given X_n, we obtain X_{n+1} by selecting any item from X_n uniformly and insert the item to the first index of the list. Notice that this means X_{n+1} could take k different permutations uniformly. Since the stochastic process is obviously a Markov chain, we have

$$H(X_{n+1} \mid X_1^n) = H(X_{n+1} \mid X_n) = \log_2 k$$

for all $n \in \mathbb{N}^+$. In the long-run, there is an equal probability of obtaining any permutation, so the stationary distribution is uniform. Therefore,

$$\lim_{n\to\infty} H(X_n) = \log_2 k!.$$

Furthermore, one may check that this Markov chain is time-invariant, irreducible and aperiodic (this is because we can always restore a permutation with 2 or 3 times of such shuffles). Now we ask the following question:

> How many steps of such shuffling process does it take such that the permutation of list obtained is independent of the initial list?

Suppose that X_n and X_0 are independent for some $n \in \mathbb{N}^+$, then necessarily $I(X_n; X_0) = 0$. This implies that the sufficient and necessary condition for such independence is

$$H(X_n \mid X_0) = H(X_n) - I(X_n; X_0) = \log_2 k!.$$

Let N_i be the index of the item selected during the i-th shuffle, then clearly X_n is completely determined by X_0 and N_1^{n-1}. Since the N_i's are independent and identically distributed, by Corollary 2.2.12 we have

$$H(X_0^n) = H(X_0, N_1^n) = H(X_0) + nH(N_1).$$

On the other hand, we also have

$$\begin{aligned} H(X_0^n) &= H(X_0) + H(X_1^n \mid X_0) \\ &= H(X_0) + H(X_n \mid X_0) + H\left(X_1^{n-1} \mid X_0, X_n\right). \end{aligned}$$

Therefore,

$$\begin{aligned} H(X_n \mid X_0) &= nH(N_1) - H\left(X_1^{n-1} \mid X_0, X_n\right) \\ &\leq nH(N_1) \\ &= n \log_2 k. \end{aligned}$$

Therefore, we necessarily need $n \log_2 k \geq \log_2 k!$, so $n \geq \frac{\log_2 k!}{\log_2 k}$. One can in fact prove further that no matter how many shuffles we perform in this manner, we still have some dependence between X_n and X_0, i.e., $H(X_n \mid X_0) < \log_2 k!$ for all $n \in \mathbb{N}^+$.

Fisher-Yates Shuffle

We have demonstrated that the naïve way of shuffling is never truly random. Now let us try to improve our algorithm. For $i = 1, 2, \cdots, k$, let M_i uniformly distributed over the index

set $\{i, i+1, \cdots, k\}$ represent the index of the item selected at random during the i-th shuffle such that the M_i-th item is then inserted to the first index of the list. Clearly, this shuffling will terminate if and only if we have performed k rounds. Furthermore, notice that by labelling each item with their index in the initial list, after the n-th round, one can reconstruct the sequence $\{M_i\}_{i=1}^{n}$ by taking the labels of items from the n-th item backwards to the first item sequentially. This means that $\{M_i\}_{i=1}^{n-1}$, and thereafter $\{X_i\}_{i=1}^{n-1}$, is completely determined by X_0 and X_n. Therefore,

$$H\left(X_1^{n-1} \mid X_0, X_n\right) = 0$$

for all $n \in \{1, 2, \cdots, k\}$. Note that since the M_i's are independent and completely determine the X_i's given X_0, by Corollary 2.2.12 we have

$$\begin{aligned} H(X_0) + H(X_k \mid X_0) + H\left(X_1^{k-1} \mid X_0, X_k\right) &= H\left(X_0^k\right) \\ &= H\left(X_0, M_1^k\right) \\ &= H(X_0) + \sum_{i=1}^{k} H(M_i). \end{aligned}$$

Therefore,

$$\begin{aligned} H(X_k \mid X_0) &= \sum_{i=1}^{k} H(M_i) - H\left(X_1^{n-1} \mid X_0, X_n\right) \\ &= \sum_{i=1}^{k} \log_2(k-i+1) \\ &= \log_2 k!. \end{aligned}$$

Thus, we have showed that this new algorithm is indeed a truly random shuffling!

3.3.2 Application: Random Walk

Consider a simple weighted undirected graph G with no loops. Let w_{ij} denote the weight of the edge $ij \in E(G)$. Suppose a moving body is traversing the graph randomly and let X_t denote its position at time t. For each $u \in V(G)$, define

$$\Pr(X_{t+1} = v \mid X_t = u) = \frac{w_{uv}}{\sum_{w \in N(u)} w_{wv}}.$$

Now we wish to investigate **how likely it is for the body to reach $v \in V(G)$ in the long-run**, i.e., given the stationary distribution π, what is $\pi(v)$ for each $v \in V(G)$?

First, we claim that

$$\pi_i = \frac{\sum_{j \in N(i)} w_{ij}}{\sum_{ij \in E(G)} w_{ij}}$$

for all $i \in V(G)$. For simplicity, let us define

$$\omega_i = \sum_{j \in N(i)} w_{ij}, \qquad \omega = \sum_{ij \in E(G)} w_{ij}$$

for all $i \in V(G)$. Let the transition probability matrix be $\boldsymbol{P}$. For each $j \in V(G)$, notice that

$$\begin{aligned}
\sum_{i \in V(G)} \pi_i P_{ij} &= \sum_{i \in V(G)} \frac{\omega_i}{\omega} \cdot \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}} \\
&= \sum_{i \in V(G)} \frac{\omega_i}{\omega} \cdot \frac{w_{ij}}{\omega_i} \\
&= \sum_{i \in V(G)} \frac{w_{ij}}{\omega} \\
&= \frac{\sum_{i \in N(j)} w_{ij}}{\omega} \\
&= \pi_j,
\end{aligned}$$

which implies that $\pi = \pi \boldsymbol{P}$. Therefore, π as defined by us previously is indeed a stationary distribution. This pretty much coincides with our intuition, because the probability of the body reaching vertex i in the long-run is given by the ratio between the total weight of the edges incident to i and that of the entire graph.

Since the process is stationary, by Proposition 3.3.12, we can compute the entropy rate as

$$\begin{aligned}
H(X) &= H(X_2 \mid X_1) \\
&= -\sum_{i \in V(G)} \pi_i \sum_{j \in V(G)} P_{ij} \log_2 P_{ij} \\
&= -\sum_{i \in V(G)} \pi_i \sum_{j \in N(i)} P_{ij} \log_2 P_{ij} \\
&= -\sum_{i \in V(G)} \frac{\omega_i}{\omega} \sum_{j \in N(i)} \frac{w_{ij}}{\omega_i} \log_2 \frac{w_{ij}}{\omega_i} \\
&= -\sum_{i,j \in V(G)^2} \frac{w_{ij}}{\omega} \log_2 \frac{w_{ij}}{\omega_i} \\
&= -\left(\sum_{i,j \in V(G)^2} \frac{w_{ij}}{\omega} \log_2 \frac{w_{ij}}{\omega} - \sum_{i,j \in V(G)^2} \frac{w_{ij}}{\omega} \log_2 \frac{w_i}{\omega} \right) \\
&= -\sum_{i,j \in V(G)^2} \frac{w_{ij}}{\omega} \log_2 \frac{w_{ij}}{\omega} + \sum_{i \in V(G)} \frac{w_i}{\omega} \log_2 \frac{w_i}{\omega},
\end{aligned}$$

where $X_1 \sim \pi$. Let (U, V) be a pair of vertices in G such that $\Pr\big((U,V) = (i,j)\big) = \frac{w_{ij}}{\omega}$, then

$$H(X) = H(U,V) - H(X_1).$$

In the case where the graph G is uniformly-weighted, one can check that the Markov chain becomes time-invariant with the stationary distribution given by

$$\pi_i = \frac{d(i)}{e(G)}.$$

Clearly, (U, V) is now uniformly distributed and so

$$H(X) = \log_2 2e(G) - H(\pi).$$

This means that during a random walk where the choice of the next vertex is uniformly distributed, the entropy rate of the walk is completely determined by the number of edges and the entropy rate of the stationary distribution.

3.3.3 Application: Hidden Markov Model

Let $\{X_i\}_{i\in\mathbb{N}^+}$ be a Markov chain and define $Y_i = \phi(X_i)$ for some function ϕ. Note that in general, $\{Y_i\}_{i\in\mathbb{N}^+}$ may not still be a Markov chain. However, it can be proved that Y is always a stationary stochastic process. By Theorem 3.3.10, the entropy rate of Y is well-defined and given by

$$H'(Y) = \lim_{n\to\infty} H\big(X_n \mid X_1^{n-1}\big),$$

where $\big\{H\big(X_n \mid X_1^{n-1}\big)\big\}_{i=1}^{\infty}$ is a non-increasing sequence. Therefore, we have

$$H'(Y) \le H\big(X_n \mid X_1^{n-1}\big)$$

for all $n \in \mathbb{N}^+$. However, if we were to approximate $H'(Y)$ with a tolerance of error of ϵ, this alone is not sufficient. Notice that by Proposition **??**,

$$H\big(Y_n \mid Y_1^{n-1}, X_1\big) \le H\big(Y_n \mid Y_1^{n-1}\big)$$

for all $n \in \mathbb{N}^+$, so

$$H\big(Y_n \mid Y_1^{n-1}, X_1\big) \le \lim_{n\to\infty} H\big(Y_n \mid Y_1^{n-1}\big) = H'(Y).$$

Therefore, we can apply some numerical methods such that

$$\big|H\big(Y_n \mid Y_1^{n-1}, X_1\big) - H\big(Y_n \mid Y_1^{n-1}\big)\big| < 2\epsilon$$

to ensure that we have a good approximation of $H'(Y)$.

Alternatively, we can also achieve this bounding condition by considering

$$\begin{aligned} I\left(Y_n; X_1 \mid Y_1^{n-1}\right) &= H\left(X_1 \mid Y_1^{n-1}\right) - H\left(X_1 \mid Y_n, Y_1^{n-1}\right) \\ &\leq H\left(X_1 \mid Y_1^{n-1}\right) \\ &\leq H\left(X_1\right). \end{aligned}$$

Furthermore, we have

$$I\left(X_1; Y_1^n\right) = H\left(X_1\right) - H\left(Y_1^n \mid X_1\right) \leq H\left(X_1\right)$$

for all $n \in \mathbb{N}^+$. This means that $\{I\left(X_1; Y_1^n\right)\}_{n=1}^{\infty}$ is a non-decreasing sequence upper bounded by $H\left(X_1\right)$, so by monotone convergence theorem,

$$\begin{aligned} H\left(X_1\right) &\geq \lim_{n\to\infty} I\left(X_1; Y_1^n\right) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} I\left(X_1; Y_i \mid Y_1^{i-1}\right). \end{aligned}$$

This means that the series

$$\sum_{i=1}^{\infty} I\left(X_1; Y_i \mid Y_1^{i-1}\right)$$

converges, and so

$$\lim_{n\to\infty} I\left(X_i; Y_n \mid Y_1^{n-1}\right) = \lim_{n\to\infty} \left|H\left(X_1\right) - H\left(Y_1^n \mid X_1\right)\right| = 0.$$

3.4 Fixed-to-Variable-Length Data Compression

Note that a fixed-to-fixed-length source code will map a sequence of n random variables to a binary string of length 2^{nR}. Now, let us consider another approach of coding: instead of encoding the entire sequence, we devise a way to encode every character in the sequence first. The code for the whole sequence then is just the concatenation of the codes of all characters.

Definition 3.4.1 ▸ Fixed-to-Variable-Length Source Code

A **Fixed-to-Variable-Length Source Code** for a random variable X is a map

$$C := \mathcal{X} \to \{0,1\}^*$$

where

$$\{0,1\}^* \bigcup_{n\in\mathbb{N}^+} \{0,1\}^n$$

is the set of all finite-length binary strings. We say that $C(x)$ is the **codeword** corresponding to x and $\ell(x)$ is the **length** of $C(x)$.

Notice that now $\ell(X)$ is a function of X, and so it is meaningful to talk about its expectation.

Definition 3.4.2 ▶ Expeted Code Length

The **expected length** $L(C)$ of a code $C: \mathcal{X} \to \{0,1\}^*$ for a random variable $X \sim p_X$ is

$$L(C) = \sum_{x\in\mathcal{X}} p_X(x)\,\ell(x) = \mathbb{E}[\ell(X)].$$

Given a string consisting of identically distributed random variables $X_1, X_2, \cdots, X_n$, we can form its codeword as $C(X_1), C(X_2), \cdots, C(X_n)$. However, the separation by comma here seems wasteful because it costs $\mathcal{O}(n)$ memory to store the commas. Can we come up with a code such that we can decode the string from the concatenation of the codewords for each character directly, without facing ambiguity? It is easy to see that a necessary condition for this is that the map C must be injective.

Definition 3.4.3 ▶ Non-singular Code

A code C is **non-singular** if C is injective.

Given a map C to encode over an alphabet, we see that C induces an encoder over all strings formed by the alphabet.

Definition 3.4.4 ▶ Extension

The **extension** of a code C is the map

$$C^*: \mathcal{X}^* \to \{0,1\}^*$$

such that $C^*(x_1x_2\cdots x_n) = C(x_1)C(x_2)\cdots C(x_n)$.

Clearly, if C^* is non-singular, then for every code, there is only on unique string which can be encoded into it.

Definition 3.4.5 ▸ Unique Decodability

A code C is **uniquely decodable** if C^* is non-singular.

It is generally hard to decide if a code is uniquely decodable. However, there is a special class for uniquely decodable codes which are relatively easy to check.

Definition 3.4.6 ▸ Prefix-Free Code

A code C is called **prefix-free** or **instantaneous** if for all $x_1, x_2 \in \mathcal{X}$ with $x_1 \neq x_2$, we have $C(x_1) \neq C(x_2)\, s$ for all $s \in \{0,1\}^*$.

For any random variable X, let $\mathcal{C}_X$ be the set of all codes, $\mathcal{S}_X$ be the set of non-singular codes, $\mathcal{U}_X$ be the set of uniquely decodable codes and $\mathcal{F}_X$ be the set of prefix-free codes on X respectively, then

$$\mathcal{F}_X \subseteq \mathcal{U}_X \subseteq \mathcal{S}_X \subseteq \mathcal{C}_X.$$

In fact, we can prove that the above inclusion is strict.

First, it is clear that $C(x) = 0$ for all $x \in \mathcal{X}$ is a singular code for X. Consider $\mathcal{X} = \{1,2,3\}$ and

$$C(x) = \begin{cases} 0 & \text{if } x = 1 \\ 010 & \text{if } x = 2 \\ 01 & \text{if } x = 3 \end{cases}.$$

Clearly, the codeword 010 is not uniquely decodable and so we have found a non-singular code which is not uniquely decodable. With some work, we can show that the code

$$C(x) = \begin{cases} 10 & \text{if } x = 1 \\ 00 & \text{if } x = 2 \\ 11 & \text{if } x = 3 \\ 110 & \text{if } x = 4 \end{cases}$$

is not prefix-free but uniquely decodable over $\{1,2,3,4\}$.

Let us study the properties of prefix-free codes in more detail.

Theorem 3.4.7 ▸ Kraft's Inequality

For any binary prefix-free code over some alphabet $\mathcal{X} := \{x_1, x_2, \cdots, x_M\}$, if $\ell_i = \ell(x_i)$ for each $x_i \in \mathcal{X}$, then

$$\sum_{i=1}^{M} 2^{-\ell_i} \leq 1.$$

Conversely, if there exists positive integers $\ell_1, \ell_2, \cdots, \ell_M$ such that

$$\sum_{i=1}^{M} 2^{-\ell_i} \leq 1,$$

then there exists a prefix-free code C over the alphabet $\{x_1, x_2, \cdots, x_M\}$ with $\ell(x_i) = \ell_i$.

Proof. Without loss of generality, let $\ell_i \leq \ell_j$ for all $i \leq j$. Let A be a complete binary tree of depth ℓ_M and order $2^{\ell_M+1} - 1$. Associate every left edge with 0 and every right edge with 1, then every code word corresponds to a unique path from the root of A to some vertex in A. Therefore, there exists an injection $f : \mathcal{X} \to V(A)$. For any $i < j$, we claim that $f(x_j)$ is not in the subtree rooted at $f(x_i)$. Suppose on contrary that $f(x_j)$ is in the subtree rooted at $f(x_i)$, then since $\ell_j \geq \ell_i$, the codeword for x_i must be a prefix to the codeword for x_j, which is a contradiction. Let A_i be the set of leaves in the subtree rooted at $v_i \in V(A)$, then $A_i \cap A_j = \varnothing$ whenever $i \neq j$. Notice that

$$|A_i| = 2^{\ell_M - \ell_i}.$$

Therefore,

$$\begin{aligned} 2^{\ell_M} &\geq \left| \bigcup_{i=1}^{M} A_i \right| \\ &= \sum_{i=1}^{M} 2^{\ell_M - \ell_i}, \end{aligned}$$

which implies that

$$\sum_{i=1}^{M} 2^{-\ell_i} \leq 1.$$

□

Remark. The converse of the theorem can be easily proven by tracing the complete binary tree backwards to re-construct the codeword for each character.

3.4.1 Optimal Code

With Theorem 3.4.7, we now have a systematic way to construct a prefix-free code. The next question is: how do we construct such a code with minimal code length? Given

$$\mathcal{X} = \{x_1, x_2, \cdots, x_M\},$$

this becomes an optimisation problem

$$\begin{aligned} &\min_{\ell_1,\ell_2,\cdots,\ell_M\in\mathbb{N}^+} \sum_{i=1}^{M} p_X(x_i)\,\ell_i \\ &\text{such that } \sum_{i=1}^{M} 2^{-\ell_i} \leq 1. \end{aligned}$$

Notice that this is an integer program. For the sake of simplicity, we may consider dropping the constraint on integrality and loosen the problem into a convex program. We can write out the Lagrangian as

$$\mathcal{L}(\boldsymbol{l},\lambda) = \sum_{i=1}^{M} p_X(x_i)\,\ell_i + \lambda\left(\sum_{i=1}^{M} 2^{-\ell_i} - 1\right).$$

At the optimum, we have

$$\frac{\partial \mathcal{L}}{\partial \ell_i} = p_X(x_i) - \lambda 2^{-\ell_i}\ln 2 = 0,$$

for all $i = 1,2,\cdots,M$. If $\boldsymbol{l}^*$ is an optimum, then $\ell_i^* \propto -\log_2 p(x_i)$ for all $i = 1,2,\cdots,M$. Notice that if $\ell_i^* = -\log_2 p(x_i)$ for all $i = 1,2,\cdots,M$, we have

$$\sum_{i=1}^{M} 2^{-\ell_i} = 1.$$

It can be proven that such a choice yields the optimal scenario, where

$$\begin{aligned} L^* &= \sum_{i=1}^{M} p_X(x_i)\,\ell_i^* \\ &= -\sum_{i=1}^{M} p_X(x_i)\log_2 p_X(x_i) \\ &= H(X). \end{aligned}$$

Clearly, the optimum of the corresponding integer program cannot be more optimal than this solution, which motivates the following theorem:

Theorem 3.4.8 ▸ Lower Bound of Expected Code Length

Let L^* be the expected code length of any prefix-free code for a random variable X, then we have $L^* \geq H(X)$, where the equality holds if and only if $2^{-\ell_i} = p_i$ for all $i = 1,2,\cdots,|\mathcal{X}|$.

Proof. Notice that

$$\begin{aligned} L^* - H(X) &= \sum_{i=1}^{|\mathcal{X}|} p_i \ell_i^* - \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{p_i} \\ &= -\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \left(2^{-\ell_i^*}\right) + \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 p_i. \end{aligned}$$

By Theorem 3.4.7, take $c := \sum_{i=1}^{|\mathcal{X}|} 2^{-\ell_i^*} \leq 1$ and define

$$r_i := \frac{2^{-\ell_i^*}}{c},$$

then clearly $\boldsymbol{r}$ is a probability vector. Observe that we can re-write

$$\begin{aligned} L^* - H(X) &= \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{p_i}{r_i} - \log_2 c \\ &= D(p \parallel r) + \log_2 \frac{1}{c} \\ &\geq 0. \end{aligned}$$

Therefore, $L^* \geq H(X)$. Clearly, the equality is achieved if and only if

$$\sum_{i=1}^{|\mathcal{X}|} 2^{-\ell_i^*} = 1$$

and $p = r$, i.e., $p_i = 2^{-\ell_i^*}$. □

Remark. If $p_i = 2^{-\ell_i^*}$ for some integer ℓ_i^*, then p_i is known as a *dyadic rational*.

Notice that to make L^* closer to $H(X)$, we need $\ell_1^* \approx \log_2 \frac{1}{p_i}$. However, based on Theorem 3.4.8 we also know it is impossible to have $\ell_1^* < \log_2 \frac{1}{p_i}$. Therefore, a natural choice of the code lengths are the ceiling of the probabilities.

Definition 3.4.9 ▸ Shannon Code

A **Shannon code** for a random variable X over $\{x_1, x_2, \cdots, x_M\}$ is a code satisfying

$$\ell_i := \ell(x_i) = \left\lceil \log_2 \frac{1}{p_i} \right\rceil.$$

We can check that for a Shannon code,

$$\begin{aligned}\sum_{i=1}^{M} 2^{-\ell_i} &= \sum_{i=1}^{M} 2^{-\left\lceil \log_2 \frac{1}{p_i} \right\rceil} \\ &\leq \sum_{i=1}^{M} 2^{-\log_2 \frac{1}{p_i}} \\ &= \sum_{i=1}^{M} p_i \\ &= 1.\end{aligned}$$

Therefore, a Shannon code can be always made prefix-free and thus uniquely decodable. Furthermore,

$$\log_2 \frac{1}{p_i} \leq \ell_i = \left\lceil \log_2 \frac{1}{p_i} \right\rceil < \log_2 \frac{1}{p_i} + 1,$$

and so

$$\begin{aligned}\sum_{i=1}^{M} p_i \log_2 \frac{1}{p_i} &\leq \sum_{i=1}^{M} p_i \ell_i \\ &< \sum_{i=1}^{M} p_i \log_2 \frac{1}{p_i} + \sum_{i=1}^{M} p_i \\ &< H(X) + 1.\end{aligned}$$

Since $\sum_{i=1}^{M} p_i \ell_i = L$, we have

$$H(X) \leq L^* < H(X) + 1.$$

Note that 1 might induce a huge interval if X is very small. To counter this, we adopt a technique known as *coding over long blocks*.

Consider any $(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ for some $n \in \mathbb{N}^+$ and define the expected codeword length per unit symbol for this vector as

$$L_n \coloneqq \frac{1}{n} \mathbb{E}\left[\ell(X_1, X_2, \cdots, X_n)\right].$$

By the previous bounds, we have

$$H(X_1^n) \leq \mathbb{E}\left[\ell(X_1, X_2, \cdots, X_n)\right] < H(X_1^n) + 1,$$

and so

$$\frac{H(X_1^n)}{n} \leq L_n^* < \frac{H(X_1^n)}{n} + \frac{1}{n}.$$

Since the source is independent and identically distributed, we have

$$H(X) \leq L_n^* < H(X) + \frac{1}{n}.$$

Therefore, the longer our block is, the smaller the difference between L_n^* and $H(X)$ will be.

Theorem 3.4.10 ▸ Wrong Code Theorem

Let $X \sim p$ and q be a probability distribution on $\mathcal{X}$. For any Shannon code with

$$\ell(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil,$$

we have

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

Lemma 3.4.11 ▸ Necessary Condition for Optimal Code

If $p_i > p_j$, then $\ell_i \leq \ell_j$.

Proof. Suppose on contrary that in an optimal code, $\ell_j < \ell_i$ but $p_i > p_j$. By exchanging the codewords for i and j to obtain a new prefix-free code, we can check that the expected codeword length strictly decreases. □

Lemma 3.4.12 ▸ Existence of Sibling Codewords

There exists an optimal code such that the codewords for the two symbols with the smallest probabilities are siblings.

Theorem 3.4.13 ▸ Optimality of Huffman Code