INFORMATION THEORY

PATRICK LEBLANC

Contents

Advisor: Liviu Nicolaescu. Started September 5, 2017. Last revised May 4, 2018.

Notation and convention

- We will denote by $|S|$ the cardinality of a finite set S.
- We denote by $\mathbb{N}$ the set of natural numbers, $\mathbb{N} = \{1, 2, \dots\}$ and we set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.
- We will use capital letters to denote random variables.
- We will use the symbol $\mathbb{E}$ when referring to expectation and the symbol $\mathbb{P}$ when referring to probability
- For any set S, contained in some ambient space X, we denote by I_S the *indicator function* of S
$$I_S : X \to \{0,1\}, \;\; I_S(x) = \begin{cases} 1, & x \in S, \\ 0, & x \notin S. \end{cases}$$
- For a discrete random variable X we will use the notation $X \sim p$ to indicate thet p is the probability mass function (pmf) of X.
- We will often refer to the range of a discrete random variable as its alphabet.

1. Introduction

Information theory is the study of how information is quantified, stored, and communicated. If we attempt to quantify, store, or communicate information via electronic means then we inevitably run into the issue of encoding — that is, how to store words, letters, and numbers as strings of binary digits. However, in doing so we might encounter a trade off between efficiency and reliability.

To illustrate this trade-off consider an example: sending a binary bit over a noisy channel. That is, whenever we send a bit there is a chance p that the bit we send will flip: e.g., if we send a 1 there is a p chance that the receiver will receive a 0. The most efficient way to send a message — say the digit 1 for sake of argument — would be to send a single 1; however, there is a possibly significant chance p that our message will be lost. We could make our message more reliable by sending 11 or 111 instead, but this vastly decreases the efficiency of the message.

Claude Shannon attacked this problem, and incidentally established the entire discipline of information theory, in his groundbreaking 1948 paper *A Mathematical Theory of Communication.* But what does information mean here? Abstractly, we can think of it as the resolving of uncertainty in a system. Before an event occurs we are unsure which of many possible outcomes could be realized; the specific realization of one of these outcomes resolves the uncertainty associated with the event.

In this paper we first formalize a measure of information: entropy. In doing so, we explore various related measures such as conditional or joint entropy, and several foundational inequalities such as Jensen's Inequality and the log-sum inequality. The first major result we cover is the Asymptotic Equipartition Property, which states the outcomes of a stochastic process can be divided into two groups: one of small cardinality and high probability and one of large cardinality and low probability. The high probability sets, known as typical sets, offer hope of escaping the trade-off described above by assigning the most efficient code to the sets with the highest-probability. We then examine similar results tor Markov Chains, which are important because important processes, e.g. English language communication, can be modeled as Markov Chains. Having examined Markov Chains, we then examine how to optimally encode messages and examine some useful applications.

2. Entropy: basic concepts and properties

2.1. **Entropy.** There are two natural ways of measuring the information conveyed in a statement: meaning and surprise. Of these meaning is perhaps the more intuitive— after all, the primary means of conveying information is to convey meaning. Measuring meaning, however, is at least as much philosophy as mathematics and difficult to do in a systemic way.

This leaves us with surprise as a measure of information. We can relate the amount of surprise in a statement to how probable that statement is; the more probable a statement the less surprising, and the less probable the more surprising. Consider the three statements below:

- The dog barked yesterday.
- The dog saved a child drowning in a well yesterday.
- QQXXJLOOCXQXIKMXQQLMQXQ

The second statement ought to be more surprising than the first. Dogs bark every day; that one should do so at some point over the course of a day is quite likely and so unsurprising. On the other hand dogs rarely save drowning children, so the occurrence of such an event is both unlikely and surprising.

So far so clear. But how might the first two statements compare with the third statement? The third statement appears to be total gibberish, and so we might be tempted to disregard it entirely. However, in another sense it is actually quite surprising. The letters Q and X do not often come up in English words, so the fact that a message was conveyed with so many Q's and X's ought to be a surprise. But how to measure this surprise?

Entropy is one such measure. A formal definition is offered below. We use the convention that $0\log(0) = 0$ based on the well known fact

$$\lim_{x\searrow 0} x\log x = 0.$$

Definition 2.1 (Entropy)**.** Fix $b > 0$ and let p be a probability distribution on a *finite or countable* set $\mathscr{X}$. We define its entropy (in base b) to be the non-negative number

$$H_b(p) := -\sum_{x\in\chi} p(x)\log_b(p(x)). \tag{2.1}$$

Let X be a discrete random variable with range (or alphabet) contained in a *finite or countable* set $\mathscr{X}$. We define then entropy of X to be the entropy of p_X, the probability mass function of X

$$H_b(X) = H_b(p_X) = -\sum_{x\in\mathscr{X}} p_X(x)\log_b p_X(x). \tag{2.2}$$

Remark 2.2. For the rest of this paper, we will use the simpler notation log when referring to $\log_2$. Also we will write $H(X)$ instead of $H_2(X)$. As the logarithm is taken base 2, the entropy is expressed in terms of bits of information. □

The entropy of a random variable X *does not depend on the values taken by* X but rather on the probability that X takes on its values. The law of subconscious statistician shows that if $X \sim p(x)$, the *expected value* of the random variable $g(X)$ is

$$\mathbb{E}_p[\,g(X)\,] = \sum_{x\in\chi} g(x)p(x). \tag{2.3}$$

We can relate the expected value of a transformation of $p(x)$ to the entropy associated with the random variable X.

Proposition 2.3. *Let X be a discrete random variable with range (or alphabet) contained in a* finite or countable *set $\mathscr{X}$ with probability mass function $p : \mathscr{X} \to \mathbb{R}$. Then,*

$$H(X) = \mathbb{E}_p\left[\log\frac{1}{p(X)}\right]. \tag{2.4}$$

Proof.

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

□

Let us return to the idea of entropy as the measure of surprise in the realization of a random variable. For any outcome $x \in \mathscr{X}$ we can think of $-\log p(x)$ as a how surprised we ought to be at observing x; note that this agrees with our intuitive conception of surprise, for the smaller the probability $p(x)$ the larger the surprise upon its observation. The entropy of p is then the amount of surprise we expect to observe when sampling $\mathscr{X}$ according to p. More properly stated, then, entropy is not a measure of surprise upon the realization of a random variable, but the amount of uncertainty present in the probability distribution of said random variable.

Obviously

$$H(X) \geq 0. \tag{2.5}$$

From the equality $\log_b(x) = \log_b(a)\log_a(x)$ we deduce immediately that

$$H_b(x) = \log_b(a)H_a(x). \tag{2.6}$$

2.2. **Joint Entropy and Conditional Entropy.** Thus far we have presented entropy only as a measure of uncertainty of the probability distribution of one random variable. For practical applications it will be useful to have a measure of uncertainty of the joint probability distribution of a sequence of random variables. This leads us to define the notions of joint and conditional entropy. We use the latter idea to derive the Chain Rule for Entropy, Theorem 2.6, which gives us another way to calculate the joint entropy of two or more random variables.

Definition 2.4 (Joint Entropy)**.** Consider a pair of discrete random variables (X, Y) with finite or countable ranges $\mathscr{X}$ and $\mathscr{Y}$, and joint probability mass function $p(x, y)$. We view the pair (X, Y) as a discrete random variable with alphabet $\mathscr{X} \times \mathscr{Y}$. The *joint entropy* $H(X, Y)$ of X, Y is the entropy of the random variable (X, Y) . More precisely entropy is

$$H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p(x,y)) = \mathbb{E}_p\left[\log\frac{1}{p(x,y)}\right]. \tag{2.7}$$

Having defined joint entropy, it is natural to define the conditional entropy of two probability distributions.

Definition 2.5 (Conditional Entropy)**.** Let (X, Y) be a pair of discrete random variables with finite or countable ranges $\mathscr{X}$ and $\mathscr{Y}$ respectively, joint probability mass function $p(x, y)$, and individual probability mass functions $p_X(x)$ and $p_Y(y)$. Then the *conditional entropy of* Y *given* X, denoted by $H(Y|X)$, is defined as

$$\begin{aligned} H(Y|X) &:= \sum_{x\in\mathscr{X}} p_X(x) H(Y|X=x) \\ &= -\sum_{x\in\mathscr{X}} p_X(x) \sum_{y\in\mathscr{Y}} p_Y(y|x)\log(p_Y(y|x)) \\ &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p_Y(y|x)) \\ &= -\mathbb{E}_p[\log(p(Y|X))]. \end{aligned} \tag{2.8}$$

Unsurprisingly, the joint and conditional entropies are intimately related to each other.

Theorem 2.6 (The Chain Rule for Entropy)**.** *Let* (X, Y) *be a pair of discrete random variables with finite or countable ranges* $\mathscr{X}$ *and* $\mathscr{Y}$ *respectively, joint probability mass function* $p(x, y)$*, and individual probability mass functions* $p_X(x)$ *and* $p_Y(y)$*. Then*

$$H(X,Y) = H(Y,X) = H(Y|X) + H(X). \tag{2.9}$$

Proof. We follow the approach in [5, p.17].

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p(x,y)) \\ &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p_X(x)p_Y(y|x)) \\ &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p_X(x)) - \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p_Y(y|x)) \\ &= -\sum_{x\in\mathscr{X}} p_X(x)\log(p_X(x)) - \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log(p_Y(y|x)) \\ &= H(X) + H(Y|X) \end{aligned}$$

□

Remark 2.7. By symmetry, we also have that $H(X,Y) = H(Y) + H(X|Y)$.

Corollary 2.8. *let (X,Y,Z) be discrete random variables with finite or countable ranges $\mathscr{X}$, $\mathscr{Y}$, and $\mathscr{Z}$ respectively. Then*

$$H(Y,X|Z) = H(Y|X,Z) + H(X|Z). \tag{2.10}$$

Proof. Follows the proof given of the theorem. □

Theorem 2.9 (Chain Rule for Entropy)**.** *Let the discrete random variables X_1,X_2,...,X_n have finite or countable ranges $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_n$, have joint mass function $p(x_1, x_2, \ldots, x_n)$, and individual probability mass functions $p_{X_1}(x), p_{X_2}(x), \ldots, p_{X_n}(x)$. Then*

$$H(X_n, X_{n-1}, \ldots, X_1) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{2.11}$$

Proof. Follows inductively from (2.9) by observing that

$$H(X_n, \ldots, X_1) = H(X_n|X_{n-1}, \ldots, X_1) + H(X_{n-1}, \ldots, X_1).$$

□

Remark 2.10. In general, $H(Y|X) \neq H(X|Y)$. However, we do have that

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Indeed,

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

This will be of use later.

2.3. Relative Entropy and Mutual Information.

Relative entropy measures the divergence between two distributions of random variables: the inefficiency of assuming that a distribution is q when in actuality it is p. If two distributions are not divergent, then we can broadly expect them to behave in the same manner; if they are divergent then they can behave in such a different manner that we cannot tell much about one distribution given the other.

Definition 2.11 (Relative Entropy)**.** The *relative entropy* or *Kullback-Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ defined on a finite or countable space $\mathscr{X}$ is

$$D(p||q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{2.12}$$

If X is a random variable with alphabet $\mathscr{X}$ and probability mass function p,

$$D(p||q) = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]. \tag{2.13}$$

We use the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$, and that $p \log \frac{p}{0} = \infty$ based on continuity arguments. Thus if there is an x such that $q(x) = 0$ and $p(x) > 0$, then $D(p||q) = \infty$.

Remark 2.12. We think of $D(p||q)$ as measuring the distance between two probability distributions; however, it is not quite a metric. While the Information Inequality (Theorem 2.21) will show that it is non-negative and zero if and only if $p = q$, it can in general fail the other criteria. Nevertheless, it is useful to think of it as measuring the distance between two distributions. □

We present one intuitive way to think of relative entropy. Let the events $A_1, A_2, \ldots, A_n$ be the possible mutually exclusive outcomes of an experiment. Assume that $q(x)$ consists of the probabilities of these outcomes, while $p(x)$ consists of the probabilities of these outcomes under different circumstances. Then

$$\left(\log \frac{1}{q_k} - \log \frac{1}{p_k} \right) \tag{2.14}$$

denotes the chance in the unexpectedness of A_k due to the different experimental circumstances. $D(p||q)$ is equal to the expected value of this change.

The interpretation of relative entropy presented above implies that two probability distributions may be related so that knowing one distribution gives us some information about the other. We shall refer to this quantity as mutual information; it measures the amount of information that one probability distribution contains about another. Mutual information be thought of as the expected change in uncertainty in one variable brought about by observing another variable.

Definition 2.13 (Mutual Information)**.** Consider two discrete random variables X, Y with ranges finite or countable ranges $\mathscr{X}$ and $\mathscr{Y}$, a joint probability mass function $p(x, y)$ and marginal probability mass functions $p_X(x)$ and $p_Y(y)$. The *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution and the product of the marginal distributions $p_X(x)p_Y(y)$.

$$\begin{aligned} I(X;Y) &:= \sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \\ &= D(p(x,y)||p_X(x)p_Y(y)) \\ &= \mathbb{E}_{p(x,y)} \log \frac{p(x,y)}{p_X(x)p_Y(y)}. \end{aligned} \tag{2.15}$$

The above definition shows that the mutual information enjoys the very important symmetry property

$$I(X,Y) = I(Y,X). \tag{2.16}$$

To gain a better understanding of the mutual information note that we can rewrite $I(X;Y)$ as

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p_X(x|y)}{p_X(x)} \\ &= -\sum_{x,y} p(x,y) \log p_X(x) + \sum_{x,y} p(x,y) \log p_X(x|y) \\ &= -\sum_{x} p_X(x) \log p_X(x) + \sum_{x,y} p(x,y) \log p_X(x|y) \\ &= H(X) - H(X|Y). \end{aligned} \tag{2.17}$$

Thus

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

We may interpret $H(X|Y)$ as the uncertainty in X given the occurrence of Y; however, this does not capture the whole of the uncertainty in the probability distribution of X. The missing piece is the mutual information $I(X,Y)$ which measures the reduction in uncertainty due to the knowledge of Y. Also, by the symmetry (2.16) we have $I(X;Y) = H(Y) - H(Y|X)$ so the mutual information between X and Y is also the reduction in uncertainty in Y due to knowledge of X. One consequence of this is that X contains as much information about Y as Y does about X.

Moreover, since $H(X,Y) = H(X) + H(Y|X)$ by (2.6), we have that

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

Finally, note that

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

so the mutual information of a variable with itself is the entropy of that variable's probability distribution. Following this logic, entropy is sometimes referred to as self-information.

The following theorem summarizes these facts.

Theorem 2.14. *Let (X,Y) two discrete random variables with with finite or countable ranges $\mathscr{X}$ and $\mathscr{Y}$, joint probability mass function $p(x,y)$, and individual probability mass functions $p_X(x)$ and $p_Y(y)$. Then we have that*

$$I(X;Y) = H(X) - H(X|Y) \tag{2.18}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{2.19}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.20}$$

$$I(X;Y) = I(Y;X) \tag{2.21}$$

$$I(X;X) = H(X) \tag{2.22}$$

We now generalize the previous results and definitions to sequences of n random variables and associated probability distributions. These results will be of use later, particularly when considering the entropy of a stochastic process.

To generalize the concept of mutual information, we will define conditional mutual information — the reduction of uncertainty of X due to knowledge of Y when Z is given — and use this quantity to prove a chain rule as we did with entropy.

Definition 2.15 (Conditional Mutual Information)**.** Let X, Y, and Z by discrete random variables with finite or countable ranges. Then the *conditional mutual information* of random variables X and Y given Z is

$$I(X,Y|Z) = H(X|Z) - H(X|Y,Z). \tag{2.23}$$

Theorem 2.16 (The Chain Rule for Information)**.** *Let X_1,X_2,...,X_n, Y be discrete random variables with finite or countable ranges. Then*

$$I(X_1, X_2, \ldots, X_n|Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1). \tag{2.24}$$

Proof. We follow [5, p.24].

$$\begin{aligned} I(X_1, X_2, \ldots, X_n|Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) && (2.25)\\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1, Y) && (2.26)\\ &= \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_n). && (2.27)\end{aligned}$$

□

2.4. **Jensen's Inequality and its Consequences.** Jensen's Inequality is a foundational inequality for convex functions. Recall that we have defined entropy to be the sum of functions of the form $-\log(\cdot)$ — as $-\log(\cdot)$ is convex, general convexity results will be useful in the study of entropy. In particular, Jensen's Inequality can be used to derive important results such as the Independence Bound on Entropy, Theorem 2.25. We begin with a definition.

Definition 2.17 (Convexity)**.** Let $I \subset \mathbb{R}$ be an interval.

(i) A function $f : I \to \mathbb{R}$ is called *convex* if for every $x_1, x_2 \in I$ and $\lambda \in [0,1]$ we have

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.28}$$

(ii) A function $f : I \to \mathbb{R}$ is *strictly convex* if we have equality in the above expression only when $\lambda \in \{0,1\}$.

(iii) A function $f : I \to \mathbb{R}$ is called *concave* if $-f$ is convex.

Let us observe that when $f : I \to \mathbb{R}$ is convex then, for any $x_1, \dots, x_N \in I$ and any $p_1, \dots, p_n \geq 0$ such that $\sum_i p_i = 1$ we have

$$f\left(\sum_i p_i x_i\right) \leq \sum_i p_i f(x_i). \tag{2.29}$$

If f is strictly convex, then above we have equality if and only if $x_1 = \cdots = x_n$.

It can be awkward to prove that functions are convex from the definition, so we shall present a sufficient criterion.

Theorem 2.18. *Let $I \subset \mathbb{R}$ be an interval, and $f : I \to \mathbb{R}$ be a function. If the function f has a second derivative that is non-negative (resp. positive) over I, then the function is convex (resp. strictly convex) over I.* □

Theorem 2.18 shows that the functions

$$x^2, \quad |x|. \quad a^x, \quad a > 1, \quad x \log x \quad (x \geq 0),$$

are convex, while the functions

$$\log x, \quad \sqrt{x}, \quad x > 0,$$

are concave.

Convexity will prove to be a useful property which is fundamental to many properties of entropy, mutual information, and other such quantities. The next result is well known.

Theorem 2.19 (Jensen's Inequality)**.** *Let $f : I \to \mathbb{R}$. If $f : I \to \mathbb{R}$ is a convex function and X is a random variable with range contained in I, then*

$$\mathbb{E}(\, f(X)\,) \geq f(\,\mathbb{E}(X)\,). \tag{2.30}$$

Moreover, if f is strictly convex and in (2.30) we have equality, then $X = \mathbb{E}(X)$ with probability 1, i.e. X is a.s. constant.

Proof. When the range of X is finite, the inequality (2.30) is the classical Jensen's inequality (2.29). The general case follows from the finite case by passing to the limit. □

Jensen's Inequality is the key result used to show several important properties of entropy.

Theorem 2.20 (Gibb's Inequality)**.** *Let $p, q : \mathscr{X} \to [0,1]$, be two probability mass functions on a finite or countable space $\mathscr{X}$. Then*

$$D(p||q) \geq 0, \tag{2.31}$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathscr{X}$.

Proof. We follow [5, p.28]. Let

$$A := \{ x \in \mathscr{X} : \ p(x) > 0 \}$$

be the support set of $p(x)$. Then

$$-D(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \overset{(2.29)}{\leq} \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \tag{$*$}$$

$$= \log \sum_{x \in A} q(x) \leq \log \sum_{x \in \mathscr{X}} q(x) = \log 1 = 0. \tag{$**$}$$

Observe that since $-\log t$ is a strictly convex function of t, we have equality sign in ($*$) if and only if the function $A \ni x \mapsto \frac{q(x)}{p(x)} \in \mathbb{R}$ is constant, i.e., there exists $c \in \mathbb{R}$ such that $q(x) = cp(x)$, $\forall x \in A$. This implies that

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

Further, we have equality in ($**$) only if

$$\sum_{x\in A} q(x) = \sum_{x\in\mathscr{X}} q(x) = 1,$$

which implies that $c = 1$. Thus, $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x. □

Corollary 2.21 (Information Inequality)**.** *Let X, Y be discrete random variables with finite or countable ranges. Then*

$$I(X;Y) \geq 0, \tag{2.32}$$

with equality if and only if X and Y are independent.

Proof. We follow [5, p.28]. Let $p(x,y)$ denote the joint probability mass function of X, Y. Then

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0,$$

with equality if and only if $p(x,y) = p(x)p(y)$. □

We may think of a "deep cause" for this result. If we find two random variables which exhibit some dependence on each other then we cannot mathematically determine which is the cause and which the effect; we can only note the degree to which they are correlated.

Corollary 2.22 (Conditioning Reduces Entropy)**.** *Let X, Y be discrete random variables with finite or countable ranges. Then*

$$H(X|Y) \leq H(X), \tag{2.33}$$

with equality if and only if X and Y are independent.

Proof.

$$0 \leq I(X;Y) = H(X) - H(X|Y) \tag{2.34}$$

□

This theorem states that knowing another random variable, Y can only reduce — and not increase — the uncertainty in X. It is important, however, to note that this is only on average. Example 2.23 below shows that it is possible for $H(X|Y = y)$ to be greater than, less than, or equal to $H(X)$, but on the whole we have $H(X|Y) = \sum_y p(y)H(X|Y = y) \leq H(X)$.

Example 2.23. Let X and Y have the following joint distribution

Y \ X	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Then $H(X) - H(\frac{1}{8}, \frac{7}{8}) = 0.544$ bit, $H(X|Y = 1) = 0$ bits, and $H(X|Y = 2) = 1$ bit. We calculate $H(X|Y) = \frac{3}{4}H(X|Y = 1) + \frac{1}{4}H(X|Y = 2) = 0.25$ bit. Thus, the uncertainty in X is increased if $Y = 2$ is observed and decreased if $Y = 1$ is observed, but uncertainty decreases on the average.

Theorem 2.24. *Let X be a discrete random variable with finite or countable range $\mathscr{X}$. Then*

$$H(X) \leq \log|\mathscr{X}|,$$

with equality if and only if X has a uniform distribution over $\mathscr{X}$.

Proof. We follow [5, p.29]. Let $u = \frac{1}{|\mathscr{X}|}$ be the uniform probability mass function over $\mathscr{X}$, and let $p(x)$ be the probability mass function for X. Then

$$D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathscr{X}| - H(X)$$

Thus by the non-negativity of relative entropy we have that

$$0 \leq D(p||u) = \log |\mathscr{X}| - H(X)$$

□

Theorem 2.25 (Independence Bound on Entropy)**.** *Let $X_1, X_2, \ldots, X_n$ be discrete random variables taking values in a finite or countable range. Then*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i), \tag{2.35}$$

with equality if and only if the X_i are independent.

Proof. We follow [5, p.30]. By the chain rule for entropies,

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

which follows from Theorem 2.22. Equality follows if and only if each X_i is independent of $X_{i-1}, \ldots, X_1$, i.e. if and only if the X_i's are independent. □

2.5. **Log Sum Inequality and its Applications.** In this section we continue to show consequences of the concavity of the logarithm function, the most important of which is the log-sum inequality. This will allow us to prove convexity and concavity for several of our measures including entropy and mutual information.

Theorem 2.26 (Log Sum Inequality)**.** *For $a_1, a_2, \ldots, a_n \in [0, \infty)$ and $b_1, b_2, \ldots, b_n \in [0, \infty)$ we have*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{2.36}$$

with equality if and only if $\frac{a_i}{b_i}$ is constant.[1]

Proof. We follow [5, p.31]. Without loss of generality, we may assume that $a_i > 0$ and $b_i > 0$. The function $f(t) = t \log t = t \log t$ is strictly convex, since $f''(t) = \frac{1}{t} \log_2 e > 0$, $\forall t \in (0, \infty)$. Hence by Jensen's inequality (2.29) we have

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i \right),$$

for $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. Setting

$$B := \sum_{j=1}^{n}, \quad \alpha_i = \frac{b_i}{B}, \quad t_i = \frac{a_i}{b_i},$$

[1] We use the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$, and $0 \log \frac{0}{0} = 0$, which follow from continuity arguments.

we obtain

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \frac{a_i}{\sum b_j},$$

which is the log-sum inequality as $\sum b_j = 1$ □

The log sum inequality will allow us to prove various convexity related results. The first will be a reproving of Theorem 2.20, which states that $D(p||q) \geq 0$ with equality if and only if $p(x) = q(x)$. By the log sum inequality we have that

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \geq \left(\sum p(x)\right) \log \frac{\sum p(x)}{\sum q(x)} = 1 \log \frac{1}{1} = 0,$$

with equality if and only if $\frac{p(x)}{q(x)} = c$. As both $p(x)$ and $q(x)$ are probability mass functions, we have that $c = 1$. Thus we have that $D(p||q) = 0$ if and only if $p(x) = q(x)$.

Armed with the log-sum inequality, we can now show that relative entropy is a convex function, and that entropy is concave.

Theorem 2.27 (Convexity of Relative Entropy)**.** *$D(p||q)$ is convex in the pair (p, q); that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions on finite or countable spaces then,*

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{2.37}$$

for $\lambda \in [0, 1]$.

Proof. We follow [5, p.32]. We can apply the log sum inequality to a term on the left hand side of (2.37) to arrive at:

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

Summing this over all values of x gives us (2.37). □

Theorem 2.28 (Concavity of Entropy)**.** *$H(p)$ is a concave function of p where p is a probability mass function on a finite or countable space.*

Proof. We follow [5, p.32]. By Theorem 2.24 we have that

$$H(p) = \log |\mathscr{X}| - D(p||u),$$

where u is the uniform distribution on $|\mathscr{X}|$ outcomes. Then, since D is convex, we have that H is concave. □

3. The Asymptotic Equipartition Property for Sequences of i.i.d. Random Variables

The *Asymptotic Equipartition Property* or *AEP* is central to information theory. In broad terms, it claims that the outcomes of certain stochastic processes can be divided into two groups: one group is small and is far more likely to occur than the other, larger, group. This property allows for efficient data compression by ensuring that some sequences generated by a stochastic process are more likely than others. In this section we show that AEP holds for sequences of i.i.d. random variables

3.1. **Asymptotic Equipartition Property Theorem.** To approach the Asymptotic Equipartition Property, we first need to cover several notions of convergence.

Definition 3.1 (Convergence of Random Variables)**.** Given a sequence of random variables $\{X_i\}$, we say that the sequence $\{X_i\}$ converges to a random variable X

(i) *in probability*, if $\forall \epsilon > 0$, $\mathbb{P}(\,|X_n - X|> \epsilon\,) \to 0$,

(ii) *in mean square*, if $E(X_n - X)^2 \to 0$,

(iii) *with probability* 1 or *almost surely* , if $\mathbb{P}\{\lim_{n\to\infty} X_n = X\} = 1$.

We can now state the Asymptotic Equipartition Theorem.

Theorem 3.2 (Asymptotic Equipartition Property). *If $\{X_i\}$ is a sequence of independent identically distributed (i.i.d.) variables, with finite or countable range $\mathscr{X}$, associated with the probability mass function $p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X), \tag{3.1}$$

in probability.

Proof. We follow [5, p. 58]. Recall that functions of independent random variables are themselves independent random variables. Then, since the X_i are i.i.d. the $\log p(X_i)$ are i.i.d. as well. Thus we have

$$\begin{aligned}-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\sum_i \log p(X_i)\\ &\to -\mathbb{E}\log p(x) \text{ in probability (by the weak law of large numbers)}\\ &= H(X).\end{aligned}$$

□

Remark 3.3. We note that if X, X' are i.i.d. discrete random variables taking values in a finite or countable space, then $\mathbb{P}(X = X') \geq 2^{-H(X)}$. To see this denote by p the common probability mass function of these two random variables and by $\mathscr{X}$ their common alphabet. Then

$$\mathbb{P}(X = X') = \sum_{x\in\mathscr{X}} \mathbb{P}(X = X' = x) = \sum_{x\in\mathscr{X}} p(x)^2 = \sum_{x\in\mathscr{X}} p(x)2^{\log p(x)} = \mathbb{E}[\,2^{\log}p(X)].$$

Using Jensen's inequality for the convex function 2^x we deduce

$$\mathbb{E}[\,2^{\log}p(X)] \geq 2^{\mathbb{E}[\log p(X)]} = 2^{-H(X)}.$$

Definition 3.4. The *ϵ-typical set* $A_\epsilon^{(n)}$, for some $\epsilon > 0$, with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathscr{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}. \tag{3.2}$$

We can use the AEP to prove several properties about the probability of sets in $A_\epsilon^{(n)}$ and the cardinality of $A_\epsilon^{(n)}$

Theorem 3.5. *If $\{X_i\}$ is a sequence of independent identically distributed (i.i.d.) variables, with finite or countable range $\mathscr{X}$, associated with the probability mass function $p(x)$, and for some $\epsilon > 0$ $A_\epsilon^{(n)}$ is the ϵ-typical set with respect to $p(x)$, then*

(i) If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$

(ii) $\mathbb{P}(A_\epsilon^{(n)}\,) > 1 - \epsilon$ for sufficiently large n.

(iii) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|S|$ denotes the cardinality of the set S.

(iv) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for sufficiently large n.

Thus the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly $2^{nH(X)}$. That is, the typical sets have high probability and are "small" if the entropy is small.

Proof. We follow [5, p.59]. The inequality (i) follows from the condition (3.2) in the definition of $A_\epsilon^{(n)}$:

$$\begin{aligned} 2^{-n(H(X)+\epsilon)} &\le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)} \\ \implies -n(H(X)+\epsilon) &\le \log p(x_1, x_2, \ldots, x_n) \le -n(H(X)-\epsilon) \\ \implies H(X) - \epsilon &\le -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \le H(X) + \epsilon \end{aligned}$$

(ii) The weak law of large numbers implies that the probability of the event $(X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. Indeed, for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(A_\epsilon^{(n)}) = \lim_{n\to\infty} \mathbb{P}\left(\left| -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) - H(X) \right| < \epsilon \right) = 1.$$

To prove (iii) we note that

$$1 = \sum_{x \in \mathscr{X}^n} p(x) \ge \sum_{x \in A_\epsilon^{(n)}} p(x) \ge 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}|,$$

and thus we have

$$|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}. \tag{3.3}$$

Finally, to prove (iv) note that, for sufficiently large n, we have that $\mathbb{P}(\ A_\epsilon^{(n)}\ ) > 1 - \epsilon$. Thus,

$$1 - \epsilon < \mathbb{P}(\ A_\epsilon^{(n)}\ ) \le \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|,$$

and thus

$$|A_\epsilon^{(n)}| \ge (1 - \epsilon) 2^{n(H(X)-\epsilon)}.$$

□

3.2. **Consequences of the AEP: Data Compression.** Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with vales in the finite alphabet $\mathscr{X}$ drawn according to the density function $p(x)$. We seek to find an efficient encoding system for such sequences of random variables. Note that, $\forall \epsilon > 0$ we can partition all sequences in $\mathscr{X}^n$ into two sets: the ϵ-typical set $A_\epsilon^{(n)}$ and its complement, $(A_\epsilon^{(n)})^c$.

Let there be a linear ordering on each of these sets, e.g. induced by lexicographic order on $\mathscr{X}^n$ defined by some linear ordering on X. Then each sequence in $A_\epsilon^{(n)}$ can be represented by giving the index of the sequence in the set. As there are at most $2^{n(H(X)+\epsilon)}$ sequences in $A_\epsilon^{(n)}$ by Theorem 3.5, the indexing ought to require no more than $n(H(X)+\epsilon)+1$ bits — the extra bit is necessary if $n(H(X)+\epsilon)$ is not an integer. These sequences are prefixed by a 0 to distinguish them from the sequences in $(A_\epsilon^{(n)})^c$, raising the total number of bits to $n(H(X)+\epsilon)+2$.

Likewise, we can index each sequence not in $(A_\epsilon^{(n)})^c$ using at most $n \log |\mathscr{X}| + 1$ bits. Adding a 1 for identification purposes increases the maximum number of bits to $n \log |\mathscr{X}| + 2$

Remark 3.6. We note the following about the above coding scheme:

- The code is one-to-one and easily decodable, assuming perfect transmission. The initial bit acts as a flag to signal the following length of code.
- We have been less efficient than we could have been, since we have used a brute force method to encode the sequences in $(A_\epsilon^{(n)})^c$. In particular, we could have taken into account the fact that the cardinality of $(A_\epsilon^{(n)})^c$ is less than the cardinality of $\mathscr{X}^n$. However, this method is still efficient enough to yield an efficient description.
- The typical sequences have short descriptions of length approximately $nH(X)$

Let x^n denote a sequence $(x_1, x_2, \dots, x_n) \in \mathscr{X}^n$ and let $\ell(x^n)$ be the length of the codeword corresponding to x^n. If n is large enough that $\mathbb{P}(\, A_\epsilon^{(n)}\,) \geq 1 - \epsilon$, the expected value of the codeword is

$$\begin{aligned}
\mathbb{E}[\,\ell(x^n)\,] &= \sum_{x^n} p(x^n)\ell(x^n) \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)\ell(x^n) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n)\ell(x^n) \\
&\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H(X)+\epsilon)+2) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n)(\,n\log|\mathscr{X}|+2\,) \\
&= \mathbb{P}(\,A_\epsilon^{(n)}\,)(\,n(H(X)+\epsilon)+2\,) + \mathbb{P}(\,(A_\epsilon^{(n)})^c\,)(n\log|\mathscr{X}|+2\,) \\
&\leq n(\,H(X)+\epsilon\,) + \epsilon n(\log|\mathscr{X}|) + 2 \\
&= n(\,H(X)+\epsilon'\,),
\end{aligned}$$

where we have that $\epsilon' = \epsilon + \epsilon\log|\mathscr{X}| + \frac{2}{n}$ can be made arbitrarily small by the correct choices of ϵ and n. We have thus proved the following result.

Theorem 3.7. *Let $X_1, \dots, X_n$ be i.i.d. random variables with finite or countable range $\mathscr{X}$ and associated with the probability mass function $p(x)$. Let $\epsilon > 0$. Then there exists a code that maps sequences x^n of length n into binary strings of length $\ell(x^n)$ such that the mapping is one-to-one (and therefore invertible) and*

$$\mathbb{E}\left[\frac{1}{n}\ell(X^n)\right] \leq H(X) + \epsilon,$$

for sufficiently large n.

3.3. High-probability Sets and the Typical Set.

We have defined $A_\epsilon^{(n)}$ so that it is a small set which contains a lot of the probability of $\mathscr{X}^n$. However, we are left with the question of whether it is the smallest such set. We will prove that the typical set has the same number of elements as the smallest set to a first order approximation in the exponent.

Definition 3.8 (Smallest Set)**.** For each $n = 1, 2, \dots$, let $B_\delta^{(n)} \subset \mathscr{X}^n$ be the smallest set such that

$$\mathbb{P}(\,B_\delta^{(n)}\,) \geq 1 - \delta.$$

We shall argue that the typical set has approximately the same cardinality as the smallest set.

Theorem 3.9. *Let $X_1, \dots, X_n$ be i.i.d. random variables with finite or countable range $\mathscr{X}$ and associated with the probability mass function $p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\mathbb{P}(\,B_\delta^{(n)}\,) > 1 - \delta$, then*

$$\frac{1}{n}\log|B_\delta^{(n)}| > H(X) - \delta' \text{ for sufficiently large } n.$$

Proof. First we prove a lemma.

Lemma 3.10. *Given any two sets A, B, such that $\mathbb{P}(A) > 1 - \epsilon_1$ and $\mathbb{P}(B) > 1 - \epsilon_2$, we have that $\mathbb{P}(A \cap B) > 1 - \epsilon_1 - \epsilon_2$.*

Proof. By the Inclusion-Exclusion Principle we have that

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \\
&> 1 - \epsilon_1 + 1 - \epsilon_2 - 1 = 1 - \epsilon_1 - \epsilon_2.
\end{aligned}$$

□

The above lemma shows that $\mathbb{P}(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 1-\epsilon-\delta$. This leads to the following sequences of inequalities.

$$\begin{aligned} 1-\epsilon-\delta \leq \mathbb{P}\Big(A_\epsilon^{(n)} \cap B_\delta^{(n)} \Big) &= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\ \leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H(X)-\epsilon)} &= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H(X)-\epsilon)} \\ &\leq |B_\delta^{(n)}| 2^{-n(H(X)-\epsilon)}. \end{aligned}$$

From here we conclude that

$$\begin{aligned} |B_\delta^{(n)}| &\geq 2^{n(H(X)-\epsilon)}(1-\epsilon-\delta) \\ \implies \log |B_\delta^{(n)}| &\geq n(H(X)-\epsilon) + \log 1 - \epsilon - \delta \\ \implies \frac{1}{n} \log |B_\delta^{(n)}| &\overset{(\delta>1/2)}{>} H(X) - \epsilon + \frac{\log \frac{1}{2} - \epsilon}{n} \\ \implies \frac{1}{n} \log |B_\delta^{(n)}| &\geq H(X) - \delta', \end{aligned}$$

for $\delta' = \epsilon - \frac{\log \frac{1}{2} - \epsilon}{n}$ which is small for appropriate ϵ and n. □

The theorem demonstrates that, to first order in the exponent, the set $B_\delta^{(n)}$, must have at least $2^{nH(X)}$ elements. However, $A_\epsilon^{(n)}$ has $2^{n(H(X) \pm \epsilon}$. Therefore, $A_\epsilon^{(n)}$ is about the same size as the smallest high-probability set.

We will now introduce notation to express equality to the first order in the exponent.

Definition 3.11 (First Order Exponential Equality)**.** Let a_n and b_n be sequences of real numbers. Then we write $a_n \doteq b_n$ if

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

Thus, $a_n \doteq b_n$ implies that a_n and b_n are equal to the first order in the exponent.

We can now restate the above results: If $\delta_n \to 0$ and $\epsilon_n \to 0$, then

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH(X)}.$$

4. Asymptotic Equipartition Property for Markov Chains

We now formally extend the results of the previous section from a sequence of *i.i.d.* random variables to a stochastic process. In particular, we will prove a version of the asymptotic equipartition property for Markov Chains. Throughout this section the random variables will be assumed valued in a finite alphabet $\mathscr{X}$.

4.1. **Background Information.** A (discrete time) stochastic process with state space $\mathscr{X}$ is a sequence $\{X_i\}_{i \in \mathbb{N}_0}$ of $\mathscr{X}$-valued random variables. We can characterize the process by the joint probability mass functions

$$p_n(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n), \;\; (x_1, \ldots, x_n) \in \mathscr{X}^n, \;\; n \in \mathbb{N}.$$

One desirable property for a stochastic process to possess is stationarity.

Definition 4.1 (Stationarity)**.** A stochastic process is called *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index. That is,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n) = \mathbb{P}(X_{1+l} = x_1, X_{2+l} = x_2 \ldots, X_{n+l} = x_n),$$

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

A famous example of a stochastic process is one in which random variable is dependent only on the random variable which precedes it and is conditionally independent of all other preceding random variables. Such a stochastic process is called a Markov Chain, and we now move to offer a formal definition.

Definition 4.2 (Markov Chain)**.** A discrete stochastic process $\{X_i\}_{i\in\mathbb{N}_0}$ with state space $\mathscr{X}$ is said to be a *Markov Chain* if, for any $n \in \mathbb{N}$, and any $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathscr{X}$, we have

$$\mathbb{P}(X_{n+1} = x_{n+1}\,|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\,) = \mathbb{P}(X_{n+1} = x_{n+1}\,|X_n = x_n\,).$$

Here we can deduce that in a Markov Chain the joint probability mass function of a sequence of random variables is

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p_1(x_2|x_1)p_2(x_3|x_2)\ldots p_n(x_n|x_{n-1}), \;\; p_k(x'|x) = \mathbb{P}(\,X_{k+1} = x'\,|X_k = x\,).$$

This motivates the importance of the next definition.

Definition 4.3 (Homogeneity)**.** A Markov Chain is called *homogeneous* if the conditional probability $p_n(x'|x)$ does not depend on n. That is, for $n \in \mathbb{N}_0$

$$\mathbb{P}(\,X_{n+1} = x'|X_n = x\,) = \mathbb{P}(\,X_2 = x'|X_1 = x\,), \;\; \forall x, x' \in \mathscr{X}.$$

Now, if $\{X_i\}_{i\in\mathbb{N}_0}$ is a Markov Chain with finite state space $\mathscr{X} = \{x_1, \ldots, x_m\}$, X_n is called the *state* at time n. A homogeneous Markov Chain is characterized by its initial state X_0 and a *probability transition matrix*

$$P = [P_{i,j}]_{1\le i,j\le m}, \;\; P_{i,j} = \mathbb{P}(\,X_1 = x_j|X_0 = x_i,\,).$$

If it is possible to transition from any state of a Markov Chain to any other with nonzero probability in a finite number of steps, then the Markov Chain is said to be *irreducible.* If the the greatest common denominator of the lengths of nonzero probability different paths from a state to itself is 1, then the Markov Chain is said to be *aperiodic.*

Further, if the probability mass function of X_n is $p_n(i) = \mathbb{P}(X_n = x_i)$, then

$$p_{n+1}(j) = \sum_{x_n} p_n(i)P_{i,j}.$$

In particular, this shows that the probability mass function of X_n is completely determined by the probability mass function of the initial state X_0.

A probability distribution π on the state space $\mathscr{X}$ is called a *stationary distribution* with respect to a homogeneous Markov chain $(X_n)_{n\in\mathbb{N}_0}$ if

$$\mathbb{P}(X_n = x) = \pi(x), \;\; \forall n \in \mathbb{N}_0, \;\; x \in \mathscr{X}.$$

If $\mathscr{X} = \{1, \ldots, m\}$ and the probability transition matrix is (P_{ij}) then we see that π is stationary if and only if

$$\pi(j) = \sum_{i=1}^{m} \pi(i)P_{i,j}, \;\; \forall j = 1, \ldots, m.$$

The next theorem describes the central fact in the theory of Markov chains. For a proof we refer to [3, 4].

Theorem 4.4. *Suppose that $\{X_n\}$ is an irreducible Markov chain with finite state space $\mathscr{X}$. Then the following hold.*

- *(i) There exists a unique stationary distribution π on $\mathscr{X}$, and for any initial data X_0 and any function $f : \mathscr{X} \to \mathbb{R}$ we have the ergodic limit*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_k) = \sum_{x\in\mathscr{X}} f(x)\pi(x), \quad \textit{almost surely.} \tag{4.1}$$

- *(ii) If additionally $\{X_n\}$ is aperiodic, then the distribution of X_n tends to the stationary distribution as $n \to \infty$, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \pi(x), \quad \forall x \in \mathscr{X}.$$

The ergodic limit (4.1) is very similar to the strong law of large numbers. We want to mention a special case of (4.1). Fix a state $x_0 \in \mathscr{X}$ and let $f : \mathscr{X} \to \mathbb{R}$ be the indicator function of the set $\{x_0\}$,

$$f(x) = \begin{cases} 1, & x = x_0, \\ 0, & x \neq x_0. \end{cases}$$

In this case, the random variable,

$$M_n(x) = \sum_{k=1}^{n} f(X_k),$$

represents the number of times the Markov chain visits the state x in the discrete time interval $\{1, \dots, n\}$. The ratio

$$F_n(x) = \frac{M_n(x)}{n}$$

then can be viewed as the frequency of these visits in this time interval. The ergodic limit then states that

$$\lim_{n\to\infty} \frac{M_n(x)}{n} = \pi(x), \quad \text{almost surely.} \tag{4.2}$$

As an example to keep in mind suppose that $\mathscr{X}$ consists of all the letters of the English alphabet (both capitalized and uncapitalized) and all the signs of punctuation. By parsing a very long English text we can produce a probability mass function π on $\mathscr{X}$ that describes the frequency of occurrence of the various symbols. We can also produce a transition probability matrix

$$P = (\, p_{xy} \,)_{x,y\in\mathscr{X}},$$

where p_{xy} describes how frequently the symbol x is followed by the symbol y.The string of symbols in the long text can be viewed as one evolution of a Markov chain with state space $\mathscr{X}$ with stationary distribution π.

4.2. **Entropy of Markov Chains.** Suppose we have a simple Markov Chain with a finite number of states $\mathscr{X} = \{x_1, x_2, \dots, x_m\}$, with a transition probability matrix $p_{i,j}$ for $i, j \in \{1, 2, \dots, m\}$ and a stationary distribution π, $\pi_i = \pi(x_i)$.

If the system is in state x_i, then it transitions to state x_k with probability $p_{i,k}$. We can regard the transition from x_i to x_k as a random variable T_i drawn according to the probability distribution $\{p_{i,1}, p_{i,2}, \dots, p_{i,n}\}$, and the entropy of X_i is

$$H(T_i) = -\sum_{k=1}^{n} p_{i,k} \log p_{i,k}.$$

The entropy of T_i will depend on i, and can be intuitively thought of as a measure of the amount of information obtained when the Markov Chain moves one step forward from the starting state of

x_i. Next, we define the average over the initial states of this quantity with respect to the stationary measure π,

$$H = \sum_{i=1}^{n} \pi_i H_i = -\sum_{i=1}^{n}\sum_{k=1}^{n} \pi_i p_{i,k} \log p_{i,k}.$$

This can be interpreted as the average amount of uncertainty removed when the Markov chain moves one step ahead. H is then defined to be the *entropy* of the Markov Chain. Moreover, H is uniquely determined by the chosen stationary probability measure π and the transition probabilities $p_{i,k}$.

We now seek to generalize these concepts to the case where we move r steps ahead for $r \in \mathbb{N}$. If the systems is in state x_i, then it is easy to calculate the probability that in the next r trials we find it in the states $x_{k_1}, x_{k_2}, \ldots, x_{k_r}$ in turn, where $k_1, k_2, \ldots, k_r$ are arbitrary numbers from 1 to n. Thus the subsequent fate of a system initially in the state A_i in the next r trials can be captured by an $\mathscr{X}^r$-valued random variable T_i^r. Moreover, the associated entropy is H_i^r and is regarded as a measure of the amount of information obtained by moving ahead r steps in the chain after starting at x_i. Then, in parallel to the above,

$$H^{(r)} = \sum_{i=1}^{n} \pi_i H_i^{(r)}.$$

This is the *r-step entropy* of the Markov Chain, which is the average amount of information given by moving ahead r steps in the chain. The one step entropy defined above can be written as $H^{(1)}$.

Since we have defined $H^{(r)}$ to be the average amount of information obtained in moving ahead r steps in a Markov Chain, then it is natural to expect that for $r, s, \in \mathbb{N}$ we have

$$H^{(r+s)} = H^{(r)} + H^{(s)},$$

or, equivalently,

$$H^{(r)} = rH^{(1)}.$$

We shall show that this later condition is true by induction on r. For $r = 1$ the definition holds trivially, as $H^{(1)} = H^{(1)}$. Now suppose that $H^{(r)} = rH^{(1)}$ for a given r. We will show that $H^{(r+1)} = (r+1)H^{(1)}$. If the system is in a state x_i, then the random variable which describes the system's behavior over the next $r+1$ transitions can be regarded as the product of two dependent variables:

(A) the variable corresponding to the transition immediately following x_i with entropy $H_i^{(1)}$ and

(B) the random variable describing the behavior of the system in the next r trials. The entropy of this variable is $H_k^{(r)}$ if the outcome of A was x_k.

Recall the relation $H(A, B) = H(A) + H(B|A)$. This gives us

$$H_i^{(r+1)} = H_i^{(1)} + \sum_{k=1}^{n} p_{i,k} H_k^{(r)}. \tag{4.3}$$

Combining this relation with the definition $H^{(r)} = \sum_{i=1}^{n} \pi_i H_i^{(r)}$ gives us

$$\begin{aligned} H^{(r+1)} &= \sum_{i=1}^{n} \pi_i H_i^{(r+1)} \\ &= \sum_{i=1}^{n} \pi_i H_i^{(1)} + \sum_{k=1}^{n} H_k^{(r)} \sum_{i=1}^{n} \pi_i p_{i,k} \\ &= H^{(1)} + \sum_{k=1}^{n} \pi_k H_k^{(r)} = H^{(1)} + H^{(r)} = (r+1)H^{(r+1)}. \end{aligned}$$

4.3. **Asymptotic Equipartition.** Consider a homogeneous Markov Chain $\{X_n\}_{n\in\mathbb{N}_0}$ with finite state space

$$\mathscr{X} = \{ x_1, \dots, x_n \},$$

and a stationary distribution $\pi : \mathscr{X} \to [0,1]$ which obeys the law of large numbers (4.2). That is, in a sufficiently long sequence of s consecutive trials the relative frequency $\frac{M_s(x)}{s}$ of the occurrence of the state $x \in \mathscr{X}$ will converge to $\pi(x)$ in probability, i.e., for any $\delta > 0$ we have

$$\lim_{s\to\infty} \mathbb{P}\left(\left|\frac{M_s(x)}{s} - \pi(x)\right| > \delta\right) = 0. \tag{4.4}$$

We set

$$P_k := \pi(x_k),$$

and we denote by p_{ij} the probability of transition from state x_i to the state x_j.

A result $\bar{x}^s \in \mathscr{X}^s$ of a sequence of s consecutive transitions in a Markov Chain be written as a sequence

$$\bar{x}^s = x_{k_1}x_{k_2}\dots x_{k_s}, \quad k_1, \dots, k_s \in \{1, 2, \dots, n\}.$$

Note that the probability of $\bar{x}^s$ materializing depends not on the part of the chain where the sequence begins due to stationarity, and is equal to

$$\mathbb{P}(\bar{x}^s) = P_{k_1}p_{k_1k_2}p_{k_2k_3}\dots p_{k_{s-1}k_s}.$$

If $i, l \in \{1, 2, \dots, n\}$, and m_{il} is the number of pairs of the form k_rk_{r+1} for $1 \le r \le s-1$ in which $k_r = i$ and $k_{r+1} = l$, then the probability of the sequence $\underline{x}^s$ can be written as

$$\mathbb{P}(\bar{x}^s) = P_{k_1}\prod_{i=1}^n\prod_{l=1}^n(p_{i,l}^{m_{il}}). \tag{4.5}$$

Theorem 4.5. *Given $\epsilon > 0$ and $\eta > 0$, there exists $S = S(\epsilon, \eta)$, sufficiently large, such that for all $s > S(\epsilon, \eta)$ the sequences $(\bar{x}^s) \in \mathscr{X}^s$ can be divided into two groups with the following properties:*

(i) The probability of any sequence $\bar{x}^s$ in the first group satisfies the inequality

$$\left|\frac{\log\frac{1}{\mathbb{P}(\bar{x}^s)}}{s} - H\right| < \eta \tag{4.6}$$

(ii) The sum of the probabilities of all sequences of the second group is less than ϵ.

That is, all sequences with the exception of a very low probability group have probabilities lying between $a^{-s(H+\eta)}$ and $a^{-s(H-\eta)}$, where a is the base of the logarithm used and H is the one-step entropy of the given chain. This parallels the partition of a probability space into typical sets and non-typical sets.

Proof. We follow the approach in [6, p.16]. With $\delta = \delta(s, \eta) >$ small enough, to be specified later, we include in the first group the sequences $\bar{x}^s \in \mathscr{X}^s$ satisfying the following two conditions

(i) $\mathbb{P}(\bar{x}^s) > 0$,

(ii) for any $i, l \in \{1, 2, \dots, n\}$ the inequality

$$|m_{i,l} - sP_ip_{i,l}| < s\delta, \quad \forall i, l \in \{1, 2, \dots, n\}. \tag{4.7}$$

holds.

All other sequences shall be assigned to group two, and we shall show that this assignment satisfies the theorem if δ is sufficiently small and s sufficiently large.

We now consider the first requirement. Suppose that $\bar{x}^s$ belongs to the first group. It follows from (4.7) that

$$m_{i,l} = sP_ip_{i,l} + s\delta\theta_{i,l} \text{ for } |\theta_{i,l}| < 1, \text{ and } i, l \in \{1, 2, \dots, n\}.$$

We now substitute the expressions for $m_{i,l}$ into (4.5). While doing so, we note that the requirement that the given sequence be a possible outcome requires $m_{i,l} = 0$ when $p_{i,l} = 0$. Thus, in (4.5) we must restrict ourselves to nonzero $p_{i,l}$, which we will denote by an asterisk on the product. Thus

$$\mathbb{P}(\bar{x}^s) = P_{k_1} \prod_i \prod_l{}^{*} (p_{i,l})^{sP_ip_{i,l}+s\delta\theta_{i,l}}$$

$$\implies \log \frac{1}{\mathbb{P}(\underline{x}^s)} = -\log P_{k_1} - s\sum_i \sum_l{}^{*} P_ip_{i,l}\log p_{i,l} - s\delta \sum_i \sum_l^{*} \theta_{i,l}\log p_{i,l}$$

$$= -\log P_{k_1} + sH - s\delta\sum_i \sum_l{}^{*} \theta_{i,l}\log p_{i,l}.$$

It follows that

$$\left|\frac{\log \frac{1}{\mathbb{P}(\bar{x}^s)}}{s} - H\right| < \frac{1}{s}\log\frac{1}{P_{k_1}} + \delta\sum_i \sum_l{}^{*} \log\frac{1}{p_{i,l}}.$$

Thus for sufficiently large s and small δ defined by

$$\eta = \delta \sum_i \sum_l{}^{*} \log\frac{1}{p_{i,l}}$$

we have that

$$\left|\frac{\log \frac{1}{\mathbb{P}(\bar{x}^s)}}{s} - H\right| < \eta.$$

Thus the first requirement of the theorem is satisfied.

We now consider the second group. For book-keeping purposes, we note that when we sum the probabilities of the sequences in this group, we exclude the probabilities of those sequences which are impossible as their probabilities are zero. Thus we are interested in the probability $P(\bar{x}^s)$ for sequences for which the inequality (4.7) fails for at least one a pair of indices $i, l \in \{1, 2, \ldots, n\}$. It suffices to calculate the quantity

$$\sum_{i=1}^{n}\sum_{l=1}^{n} \mathbb{P}(|m_{i,l} - sP_ip_{i,l}|\geq s\delta).$$

Fix the indices i and l. By (4.4) we have that sufficiently large s

$$P\left(|m_i - sP_i|< \frac{\delta}{2}s\right) > 1-\epsilon.$$

Then if the inequality $|m_i - sP_i|< \frac{\delta}{2}s$ is satisfied, and if we make s is sufficiently large, then m_i can be made arbitrarily large. Thinking of the transition $i \to l$ as a Bernoulli trial with success probability p_{il} we expect that during m_i trials we would observe m_ip_{il} such transitions. The weak law of large numbers then shows that if s is sufficiently large and thus m_i is sufficiently large we have

$$\mathbb{P}\left(\left|\frac{m_{i,l}}{m_i} - p_{i,l}\right| < \frac{\delta}{2}\right) > 1-\epsilon.$$

Combining these results gives us that the probability of satisfying the inequalities

$$|m_i - sP_i|< \frac{\delta}{2}s \tag{4.8}$$

$$|m_{i,l} - m_ip_{i,l}|< \frac{\delta}{2}m_i \leq \frac{\delta}{2}s \tag{4.9}$$

exceeds $(1-\epsilon)^2 > 1-2\epsilon$. However, it follows from (4.8) that

$$|p_{i,l}m_i - sP_ip_{i,l}|< p_{i,l}\frac{\delta}{2}s < \frac{\delta}{2}s,$$

which together with (4.9) and the triangle inequality gives us

$$|m_{i,l} - sP_ip_{i,l}|< \delta s. \tag{4.10}$$

Thus for any i and l, we have that for sufficiently large s

$$\mathbb{P}\Big(|m_{i,l} - sP_ip_{i,l}|< \delta s \Big) > 1-2\epsilon,$$

which implies that

$$\mathbb{P}\Big(|m_{i,l} - sP_ip_{i,l}|> \delta s \Big) < 2\epsilon,$$

and from this it follows that

$$\sum_{i=1}^{n}\sum_{l=1}^{n}\mathbb{P}\Big(|m_{i,l} - sP_ip_{i,l}|> \delta s \Big) < 2n^2\epsilon.$$

As the right hand side of the above inequality can be made arbitrarily small as ϵ becomes arbitrarily small, we have that the sum of the probabilities of all the sequences in the second group can be made arbitrarily small for sufficiently large s. □

In the remainder of this section we will assume that the base of the logarithms employed is $a = 2$. We denote by $\mathscr{X}^s_{\epsilon,\eta}$ the collection of sequences in the first group and we will refer to such sequences as (ϵ,η)-typical. By construction

$$\mathbb{P}(\,\mathscr{X}^s_{\epsilon,\eta}\,) > 1-\epsilon$$

if s is sufficiently large. Now observe that

$$1 \geq \sum_{\bar{x}^s\in\mathscr{X}^s_{\epsilon,\eta}} \mathbb{P}(\bar{x}^s) \geq 2^{-s(H+\eta)}|\mathscr{X}^s_{\epsilon,\eta}|,$$

and we conclude

$$|\mathscr{X}^s_{\epsilon,\eta}| \leq 2^{s(H+\eta)}.$$

Finally observe that

$$(1-\epsilon) < \mathbb{P}(\,\mathscr{X}^s_{\epsilon,\eta}\,) = \sum_{\bar{x}^s\in\mathscr{X}^s_{\epsilon,\eta}} \mathbb{P}(\bar{x}^s) \leq |\mathscr{X}^s_{\epsilon,\eta}|2^{-(H-\eta)},$$

so that

$$|\mathscr{X}^s_{\epsilon,\eta}| > (1-\epsilon)2^{(H-\eta)}.$$

We see that the (ϵ,η)-typical sets $\mathscr{X}^s_{\epsilon,\eta}$ satisfy all the properties of the ϵ-typical sets in the Asymptotic Equipartition Theorem 3.5. Arguing as in the proof of Theorem 3.7 we deduce the following result.

Theorem 4.6. *Consider a homogeneous ergodic Markov chain $(X_n)_{n\geq 1}$ on the finite set $\mathscr{X}$ with stationary distribution π and associated entropy H. Fix $\epsilon > 0$ and denote by B the set of finite strings of* 0*-s and* 1*-s. Then, for every $s > 0$ sufficiently large there exists an injection $c_s : \mathscr{X}^s \to B$ such that if $\ell_s(\bar{x}^s)$ denotes the length of the string $c_s(\bar{x}^s)$ we have*

$$\mathbb{E}\left[\frac{1}{s}\ell_s(\bar{X}^s)\right] \leq H+\epsilon,$$

where $\bar{X}^s$ is the random vector $(X_1,\dots,X_s)$.

5. Coding and Data Compression

We now begin to establish the limit for information compression by assigning short descriptions to the most probable messages, and assigning longer descriptions to less likely messages. In this chapter we will find the shortest average description length of a random variable by examining its entropy.

5.1. **Examples of Codes.**

In the sequel $\mathscr{X}$ will denote a finite set called alphabet. It will equipped with a probability distribution p. We will often refer to the elements of $\mathscr{X}$ as the symbols or the letters of the alphabet.

Think for example that $\mathscr{X}$ consists of all the letters (small and large caps) of the English alphabet together with all the sign of punctuation and a symbol for a blank space (separating words). One can produce a probability distribution on such a set by analyzing a very long text an measuring the frequency with each each of these symbols appears in this text.

Definition 5.1 (Codes)**.** Suppose that $\mathscr{X}$ and $\mathscr{D}$ are finite sets.

(i) We denote by $\mathscr{D}^*$ the collection of words with alphabet $\mathscr{D}$,

$$\mathscr{D}^* := \bigcup_{n\in\mathbb{N}} \mathscr{D}^n.$$

The length of a word $w \in \mathscr{D}^*$ is the natural number n such that $w \in \mathscr{D}^n$.

(ii) A *source code* C or simply *code* for $\mathscr{X}$ based on $\mathscr{D}$ is a mapping

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

For $x \in \mathscr{X}$ we denote by $\ell_C(x)$ the length of $C(x)$.

Example 5.2. Suppose $\mathscr{X} = \{\text{yes, no}\}$ and $\mathscr{D} = \{0, 1\}$. Then $C(\text{yes}) = 00$ and $C(\text{no}) = 11$ is a source code.

Definition 5.3 (Expected Length)**.** Let p be a probability mass function on the finite alphabet $\mathscr{X}$, and let $\mathscr{D}$ be a finite set. The *expected length* $L(C)$ of a source code $C : \mathscr{X} \to \mathscr{D}^*$ is the expectation of the random variable $\ell_C : \mathscr{X} \to \mathbb{N}$, i.e.,

$$L(C) = \mathbb{E}_p[\,\ell_C\,] = \sum_{x\in\mathscr{X}} p(x)\ell_C(x).$$

Without loss of generality, we may assume that the D-ary alphabet is $\mathscr{D} = \{0, 1, \dots, D-1\}$. Now we consider some simple examples.

Example 5.4. Let $\mathscr{X} = \{1, 2, 3, 4\}$, $\mathscr{D} = \{0, 1\}$. Consider the following probability distribution and code $C : \mathscr{X} \to \mathscr{D}^*$

$$p(1) = \frac{1}{2},\;\; C(1) = 0,\;\; p(2) = \frac{1}{4},\;\; C(2) = 10,$$

$$p(3) = \frac{1}{8},\;\; C(3) = 110,\;\; p(4) = \frac{1}{8},\;\; C(4) = 111.$$

The entropy $H(p) = 1.75$ bits, and the expected length is $L(C) = 1.75$ bits as well — thus we have an encoding scheme with the same average length as the entropy. Moreover, any sequence of bits can be uniquely decoded into a sequence of symbols of $\mathscr{X}$. Later on we will see that the entropy is in fact a lower bound on the expected length of the code.

Now we shall define more stringent conditions on codes. Let x^n denote $(x_1, x_2, \dots, x_n)$.

Definition 5.5 (Nonsingular)**.** Let $\mathscr{X}$ and $\mathscr{D}$ be finite sets. A code $C : \mathscr{X} \to \mathscr{D}^*$ is said to be *nonsingular* if it is injective. That is, different elements of $\mathscr{X}$ are encoded by different strings in $\mathscr{D}^*$,

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

If an encoding scheme is nonsingular, then we have an unambiguous description of a single letter of the alphabet $\mathscr{X}$. However, we are not guaranteed an unambiguous description if we send a string of letters. One possible method to ensure a clear decoding scheme is to insert a special character, such as a comma, in between every word. However, this is inefficient. Rather, we shall develop the idea of self-punctuating or instantaneous codes. To do this, we will first define the extension of a code.

Definition 5.6 (Extension)**.** Let $\mathscr{X}$ and $\mathscr{D}$ be finite sets. Then the *extension* C^* of a code $C : \mathscr{X} \to \mathscr{D}^*$ is the mapping

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

defined by

$$C(x_1x_2 \ldots x_n) = C(x_1)C(x_2) \ldots C(x_n),\ \forall n \in \mathbb{N},\ \ x_1, \ldots, x_n \in \mathscr{X}.$$

where $C(x_1)C(x_2) \ldots C(x_n)$ indicates the concatenation of the relevant codewords.

Example 5.7. If $C(x_1) = 00$ and $C(x_2) = 11$, then $C(x_1x_2) = 0011$.

Definition 5.8 (Uniquely Decodable)**.** A code is *uniquely decodable* if its extension is non-singular.

If a code is uniquely decodable, then any encoded string has only one possible source string corresponding to it. However, the source string may be unintelligible without first decoding the entire encoded string.

Definition 5.9 (Instantaneous Code)**.** A code is called a *prefix code* or a *instantaneous code* if no codeword is a prefix of any other codeword.

An instantaneous code can be decoded without reference to future or past codewords. In an instantaneous code, the symbol x_i can be decoded as we arrive at the end of a codeword containing it. The code is *self-punctuating* — we can look at any sequence of encoded symbols and separate them into the codewords. For instance, the encoding scheme,

$$C(1) = 0,\ \ C(2) = 10,\ \ C(3) = 110,\ \ C(4) = 111,$$

covered in the Example 5.4 is a self-punctuating code. Consider the sequence 01011111010. This uniquely breaks down into the codewords $0, 10, 111, 110, 10$.

5.2. **Kraft Inequality.** Suppose we are given a probability mass function p on the finite alphabet $\mathscr{X}$. For simplicity we assume

$$\mathscr{X} = \{1, \ldots, m\}.$$

Our goal is to construct instantaneous codes of minimum expected length to describe a given source. There are, however, restrictions — we cannot assign short codewords to all source symbols while creating a prefix free code. The sets of codeword lengths possible for instantaneous codes are constrained by the Kraft's Inequality.

Theorem 5.10 (Kraft Inequality)**.** *Let $\mathscr{D}$ be a finite set and $\mathscr{X}$ a finite alphabet. Suppose that $C : \mathscr{X} \to \mathscr{D}^*$ is an instantaneous code (prefix code) over an alphabet of size D. Set*

$$l_i = \ell_C(i),\ \ i = 1, \ldots, m.$$

Then the codeword lengths $l_1, l_2, \ldots, l_m$ must satisfy the inequality

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

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

Proof. We follow the ideas in [5, §5.2]. Consider a D-ary tree in which each node has D children. Let the branches of the tree represent the symbols of the codeword. For example, the D branches arising from the root node represent the D possible values of the first symbol of the codeword. Then each codeword is represented by a node on this tree: the path from the root traces out the symbols of the codeword. The prefix condition on the codewords implies that no codeword is an ancestor of any other codeword on the tree. Hence, each codeword eliminates its descendants as possible codewords.

Let l_{max} be the length of the longest codeword of the set of codewords. Consider all nodes of the tree at level l_{max}. Some of them are codewords, and some are descendants of codewords, and some are neither. A codeword at level l^i has $D^{l_{max}-l_i}$ descendants at level l_{max}. Crucially, these descendant sets *are disjoint*. The total number of nodes in these sets must be less than or equal to $D^{l_{max}}$. Hence summing over all the codewords gives us

$$\sum D^{l_{max}-l_i} \leq D^{l_{max}} \text{ or, equivalently, } \sum D^{-l_i} \leq 1.$$

This is Kraft's Inequality.

Conversely, given any set of codeword lengths $l_1, l_2, \ldots, l_m$, that satisfy the Kraft Inequality, we can always construct a tree such as the one constructed above. Label the first node (lexicographically) of depth l_1 as the codeword 1 and cross off its descendants. Label the first remaining node of depth l_2 as codeword 2, and so on. Proceeding this way constructs a prefix code with the specified $l_1, l_2, \ldots, l_m$. □

5.3. **Optimal Codes.** We have shown that any codeword which satisfies the prefix condition must satisfy the Kraft inequality, and that the Kraft inequality is a sufficient condition for the existence of a codeword set with the specified set of codeword lengths. Now we consider a probability mass function

$$p : \mathscr{X} = \{1, \ldots, m\} \to [0,1], \ \ p(i) = p_i,$$

and we seek a prefix code with the minimum expected length.

By the above, this is equivalent to finding the set of lengths $l_1, l_2, \ldots, l_m$ satisfying the Kraft inequality and whose expected length is less than the expected length of any other prefix code.This is a standard optimization problem: minimize

$$L = \sum p_i l_i$$

over all integers $l_1, l_2, \ldots, l_m$ satisfying

$$\sum D^{-l_i} \leq 1.$$

Let us first observe that a minimizer satisfies the more stringent constraint

$$g = \sum_{i=1}^{m} D^{-x_i} = 1.$$

Indeed, if a minimizer was located in the open set $\{g < 1\}$ then the gradient of L at that point would have to be zero.On the other hand, the gradient of L is the nonzero vector $(p_1, \ldots, p_m)$.

We will utilize Lagrange multipliers to calculate the form of the minimizing lengths l_i^*. That is, we will minimize the function $L = \sum p_i l_i$ subject to the constraint $g = \sum D^{-l_i} = 1$. Thus we have

$$\nabla L = \lambda \nabla D g$$
$$\sum D^{-l_i} = 1.$$

Differentiating with respect to the l_i gives us

$$p_i = -\lambda D^{-l_i} \ln D.$$

We deduce

$$D^{-l_i} = -\frac{p_i}{\lambda \ln D}, \quad 1 = \sum_i D^{-l_i} = -\frac{1}{\lambda \ln D}\sum_i p_i = -\frac{1}{\lambda \ln D}.$$

We find that $\lambda \ln D = -1$ and

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

which in turn yields

$$l_i^* = -\log_D p_i.$$

This non-integer choice of codeword length yields an expected codeword length of

$$L^* = \sum p_i l_i^* = -\sum p_i \log_D p_i = H_D(p).$$

It shows that the expected length of an instantaneous code is at least $H_D(p)$. However, since the l_i must be integers, it is not always possible to attain the values listed above; rather, we must choose a set of codeword lengths l_i “close” to the optimal set.

We want to give an alternate proof of this lower bound

Theorem 5.11. *For any probability mass function p on the alphabet $\mathscr{X} = \{1, \dots, m\}$, the expected length $L(C)$ of any instantaneous D-ary code C on $\mathscr{X}$ is not smaller than the entropy of p, that is*

$$L(C) \geq H_D(p),$$

with equality if and only if $D^{-l_i} = p_i$.

Proof. We follow the ideas in the proof of [5, Thm. 5.3.1]. Consider the difference between the expected length of a code and the entropy

$$L - H_D(p) = \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} = -\sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i.$$

Now if we set

$$r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}} \text{ and } c := \sum D^{-l_i},$$

we have

$$L - H = \sum p_i \log_D \frac{p_i}{r_i} - \log_D c = D(p||r) + \log_D \frac{1}{c} \geq 0,$$

by the non-negativity of relative entropy, and the fact that $c \leq 1$ by the Kraft Inequality. Thus $L \geq H$ with equality if and only if $p_i = D^{-l_i}$, that is if and only if $-\log_D p_i$ is an integer for all i. □

We need to introduce some convenient terminology.

Definition 5.12 (D-adic distributions)**.** A probability mass function p on a finite set $\mathscr{X}$ is called *D-adic* if there exists a constant $c > 0$ and a function $l : \mathscr{X} \to \mathbb{N}$ such that

$$p(x) = cD^{-l(x)}, \ \forall x \in \mathscr{X}.$$

Thus, we have equality in the above theorem if and only if the distribution p is D-adic. The proof also suggests a procedure for finding an optimal code with respect to a given probability mass function p on $\mathscr{X}$: find the D-adic distribution possessing the lowest relative entropy with respect to p. This distribution provides the set of code-word lengths, and then we can construct the code by following the algorithm outlined in the proof of the Kraft inequality.

In practice, this methodology may not be trivial to apply as it is not obvious how to find a D-adic distribution with minimal relative entropy. This will be addressed in the section on Huffman coding.

5.4. **Bounds on the Optimal Code Length.** Suppose that $\mathscr{X} = \{1, \dots, m\}$ is an alphabet equipped with a probability mass function p. We will produce a code that achieves an expected description length L within 1 bit of the lower bound, i.e., $H(p) \leq L \leq H(p) + 1$. We follow [5, p.113].

We seek a D-adic probability distribution r minimizing

$$L = D(p||r) + \log_D \frac{1}{c}.$$

The choice of word lengths $l_i = \log_D \frac{1}{p_i}$ yields $L = H$. Since $\log_D \frac{1}{p_i}$ may not be an integer, the natural thing to do is to round up,

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

where $\lceil x \rceil$ is the smallest integer greater than or equal to x. Note that these lengths satisfy the Kraft inequality as

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

This shows that there exists a prefix code $C : \mathscr{X} \to \mathscr{D}^*$ with code lengths $\ell_C(i) = l_i$ and this code satisfies

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

Multiplying by p_i and summing over i gives us

$$H_D(p) \leq L = \mathbb{E}_p[\ell_C] \leq H_D(p) + 1.$$

A prefix code code with optimal mean length $L_{\min}(p)$ can only improve on this code so $L_{\min}(p) \leq L$. On the other hand, by Theorem 5.11 we have $L_{\min}(p) \geq H_D(p)$. We have thus proved the following theorem.

Theorem 5.13. *Let $L_{\min}(p)$ denote the expected length of an optimal instantaneous code the associated expected length of an optimal code $C_{\min} : \mathscr{X} = \{1, \dots, m\} \to \mathscr{D} = \{1, \dots, D\}$ with respect to a probability mass function p on a finite set $\mathscr{X}$. Then*

$$H_D(p) \leq L_{\min}(p) < H_D(p) + 1. \tag{5.1}$$

Definition 5.14. Let $\mathscr{X} = \{1, \dots, m\}$ be a finite alphabet, $\mathscr{D} = \{1, \dots, D\}$ and p a probability mass function on $\mathscr{X}$, $p_i := p(i)$. A code $C : \mathscr{X} \to \mathscr{D}^*$ such that

$$\ell_C(i) = \left\lceil \log_D \frac{1}{p_i} \right\rceil$$

is called a *Shannon code.*

We see that Shannon codes are nearly optimal. In the above theorem, the upper error bound is 1 bit because $\log \frac{1}{p_i}$ is not always an integer. We can reduce this error per symbol by spreading it out over many symbols. Consider a system in which we send a sequence of n symbols from $\mathscr{X}$, which interpret as a sequence of independent $\mathscr{X}$-valued random variables $X_1, \dots, X_n$ each distributed as p. We can consider this string of symbols to be a super-symbol from the alphabet $\mathscr{X}^n$ equipped with the probability distribution $p^{\otimes n}$.

For simplicity assume $D = 2$ and we are given an optimal prefix code

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

A word $\bar{x}^n = (x_1, \dots, x_n)$ is encoded by the word $C(x_1) \cdots C(x_n)$ and has length

$$\ell_C(x_1, \dots, x_n) = \ell_C(x_1) + \cdots + \ell_C(x_n).$$

Define

$$L_n(C) := \frac{1}{n} \sum p(x_1) \cdots p(x_n) \ell_C(x_1, \dots, x_n) = \frac{1}{n} \mathbb{E}_p[\ell_C(x_1, \dots, x_n)].$$

The quantity L_n can be interpreted as the average cod length per letter in messages encoding words of length n in the alphabet $\mathscr{X}$. If we apply the bounds derived above we arrive at

$$H(X_1,\ldots,X_n) \leq \mathbb{E}_p[\ell_C(X_1,\ldots,X_n)] < H(X_1,\ldots,X_n)+1.$$

As the X_i are i.i.d. we have that $H(X_1,\ldots,X_n) = \sum H(X_i) = nH(p)$. Then dividing the above by n gives us

$$H(p) \leq L_n(C) < H(p) + \frac{1}{n}.$$

Thus, by using large block lengths we can achieve an expected code-length per symbol arbitrarily close to the entropy.

We now consider a natural question — what happens to the expected description length if the code is designed for the wrong distribution? The wrong distribution might arise if as the best estimate we can make of an unknown true distribution: the Shannon code assignment $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ designed for the probability mass function $q(x)$. Here we will not achieve the expected length $L \approx H(p) = -\sum p(x)\log p(x)$. Further, the increase in expected description length due to this incorrect distribution is the relative entropy $D(p||q)$, and thus we can think of $D(p||q)$ as the increase in descriptive complexity due to incorrect information.

Theorem 5.15 (Wrong Code Theorem). *Let $\mathscr{X}$ be a finite set and let p be a probability mass function on $\mathscr{X}$. Then the expected length under p of the code assignment $\ell(x) = \lceil \log \frac{1}{q(x)} \rceil$ satisfies*

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

Proof. We follow [5, p.115]. The expected codeword length is

$$\mathbb{E}_p[l(X)] = \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil < \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right)$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 = D(p||q) + H(p) + 1.$$

The lower bound is derived similarly. □

Thus assuming that the distribution is $q(x)$ when it is actually $p(x)$ incurs a penalty of $D(p||q)$ in the average description length.

5.5. **Kraft Inequality for Uniquely Decodable Codes.** We have shown that any instantaneous code must satisfy the Kraft inequality. However, the class of uniquely decodable codes is larger than the class of instantaneous codes, so one expects to achieve a lower expected codeword length if L is minimized over all uniquely decodable codes. Here we seek to prove that the class of uniquely decodable codes does not offer any further possibilities for the set of codeword lengths than do instantaneous codes, and will follow Karush's proof of the following theorem:

Theorem 5.16 (McMillan). *The codeword lengths of any uniquely decodable D-ary code must satisfy the Kraft inequality*

$$\sum D^{-l_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.

Proof. We follow [5, p.116]. Consider C^k, the k^{th} extension of the code — that is, the code formed by the concatenation of k repetitions of the given uniquely decodable code C. By the definition of unique decodability, the k^{th} extension of the code is nonsingular. As there are only D^n different D-ary strings of length n, unique decodability implies that the number of code sequences of length n in the k^{th} extension of the code must be no greater than D^n, and we use this oservation to prove the Kraft inequality.

Let the codeword lengths of the symbols $x \in \mathscr{X}$ be denoted by $l(x)$. For the extension code, the length of the code sequence is

$$l(x_1, \ldots, x_n) = \sum_{i=1}^{k} l(x_i)$$

The inequality that we wish to prove is $\sum D^{-l_i} \leq 1$, and the trick is to consider the k^{th} power of this quantity. Thus,

$$\begin{aligned}\left(\sum_{x \in \mathscr{X}} D^{-l(x)}\right)^k &= \sum_{x_1 \in \mathscr{X}} \sum_{x_2 \in \mathscr{X}} \cdots \sum_{x_k \in \mathscr{X}} D^{-l(x_1)} D^{-l(x_2)} \ldots D^{-l(x_n)} \\ &= \sum_{x_1, x_2, \ldots, x_k \in \mathscr{X}^k} D^{-l(x_1)} D^{-l(x_2)} \ldots D^{-l(x_n)} \\ &= \sum_{x^k \in \mathscr{X}^k} D^{-l(x^k)}.\end{aligned}$$

If we now gather the terms by word length we obtain

$$\sum_{x^k \in \mathscr{X}^k} D^{-l(x^k)} = \sum_{m=1}^{kl_{max}} a(m) D^{-m},$$

where l_{max} is the maximum codeword length and $a(m)$ is the number of source sequences x^k mapping into codewords of length m. But the code is uniquely decodable, so there is at most one sequence mapping into each code m-sequence and so there are at most D^m code m-sequences. Thus, $a(M) \leq D^m$ and so we have

$$\begin{aligned}\left(\sum_{x \in \mathscr{X}} D^{-l(x)}\right)^k &= \sum_{m=1}^{kl_{max}} a(m) D^{-m} \\ &\leq \sum_{m=1}^{kl_{max}} D^m D^{-m} \\ &= kl_{max},\end{aligned}$$

and thus we can deduce that

$$\sum_j D^{-l_j} \leq (kl_{max})^{\frac{1}{k}}.$$

Since this inequality is true for all k, it is true in the limit as $k \to \infty$. Since $(kl_{max})^{\frac{1}{k}} \to 1$, we have

$$\sum_j D^{-l_j} \leq 1,$$

which is the Kraft inequality.

Conversely, given any set of $l_1, l_2, \ldots, l_m$ satisfying the Kraft inequality, we can construct an instantaneous code as in the proof of the Kraft inequality. Since every instantaneous code is uniquely decodable, we have also constructed a uniquely decodable code. □

This theorem implies the counter-intuitive result that the class of uniquely decodable codes does not offer any further choices for the set of codeword lengths than does the class of prefix codes. That is, the set of achievable codeword lengths is the same for uniquely decodable and instantaneous codes. Thus, the bounds derived on the optimal codeword lengths continue to hold even when we expand the class of allowed codes to the class of all uniquely decodable codes.

5.6. **Huffman Codes.** We want to describe an important class of optimal prefix or instantaneous codes. These are the so called *Huffman codes.* For simplicity we will work will work with a binary alphabet, $D = 2$. The Huffman codes are constructed inductively.

Here it is important to note that the optimal codes are non-unique — inverting all the bits, or exchanging codewords of the same length, will give another optimal code. We need to introduce some notation. For any probability mass function p on an alphabet $\mathscr{X}$ we denote by $L_{\min}(p)$ the expected length (with respect to p) of an optimal code.

We order the letters $\{x_1, \dots, x_n\}$ of our alphabet $\mathscr{X}$ so that the associated probability masses are ordered, $p_1 \geq p_2 \geq \dots \geq p_m$. As usual we set

$$l_i := \ell_C(x_i).$$

Recall that a code is optimal if $\mathbb{E}_p[\ell_C] = \sum p_i l_i$ is minimal among all binary instantaneous codes $\mathscr{X} \to \{0,1\}^*$.

First we show that among the optimal prefix codes there are some with special properties.

Lemma 5.17. *For any probability mass function p on a finite set $\mathscr{X}$, there exists an optimal instantaneous code (with minimum expected length) that satisfies the following properties:*

(i) The lengths are ordered inversely with the probabilities, i.e., if $p_j > p_k$, then $l_j \leq l_k$.
(ii) The two longest codewords have the same length.
(iii) The two longest codewords differ only in the last bit and correspond to the two least likely symbols.

Proof. We follow the proof of [5, Lemma 5.8.1]. Consider an optimal prefix code $C : \mathscr{X} \to \{0,1\}^*$, $l_i := \ell_C(x_i)$. We will show that we can modify C so it becomes an optimal prefix code satisfying the properties (i)-(iii) above.

Step 1. The optimal code C satisfies (i). More precisely, we will show that if $p_j > p_k$ then $l_j \leq \ell_k$.

We argue by contradiction. Assume that $l_j > l_k$. We modify C by swapping the codewords for x_j and x_k. More precisely, we construct a new code C' with the codewords of j and k of C interchanged,

$$C'(x_j) = C(x_k), \;\; C'(x_k) := C(x_j).$$

Set $l'_i := \ell_{C'}(x_i)$. Then

$$L(C') - L(C) = \sum p_i l'_i - \sum p_i l_i = p_j l_k + p_k l_j - p_j l_j - p_k l_k = (p_j - p_k)(l_k - l_j) < 0.$$

This contradicts the fact that C is optimal.

Step 2. We show that the optimal code C satisfies (ii). Again, we argue by contradiction. If the two longest codewords are not of the same length, one can delete the last bit of the longer one, preserving the prefix property and achieving a lower expected codeword length. Hence, the two longest codewords must have the same length. By the first property, the longest codewords correspond to the least probable symbols.

Step 3. We say that two codewords are *siblings* if they coincide except for the last bit. E.g.,the codewords 1011 and 1010 are siblings. We will show that we can modify C to another optimal instantaneous code C' so that two longest of the longest codewords are siblings and correspond to the two least likely symbols.

If there is a maximal-length codeword without a sibling, we can delete the last bit of the codeword and still satisfy the prefix property. This reduces the average codeword length and contradicts the optimality of the code. Hence, every maximal length codeword has a sibling.

Suppose that the maximal lengths codes correspond to the letters $y_1, y_2, \dots, y_{2k-1}, y_{2k} \in \mathscr{X}$ where $C(y_{2i-1})$ and $C(y_{2i})$ are siblings. Now choose a permutation ϕ of $\{1, \dots, 2k\}$ such that

$$p(y_{\phi(1)}) \geq \cdots \geq p(y_{\phi(2k)}).$$

and modify C to a new code C' such that

$$C'(y_{\phi(i)}) = C(y_i), \;\; C'(x) = C(x), \;\; \forall x \neq y_1, \dots, y_{2k}.$$

Note that the code words $C'(y_{\phi(2k-1)})$ and $C'(y_{\phi(2k)})$ have maximal length, are siblings, and correspond to the least two probable letters with maximal code length.

□

Definition 5.18 (Canonical Codes)**.** An optimal instantaneous code satisfying the properties of the lemma is called a *canonical code.*

To any probability mass function $\vec{p} = (p_1, p_2, \dots, p_m)$ on the alphabet $\{1, 2, \dots, m\}$ alphabet of size m, with $p_1 \geq p_2 \geq \dots \geq p_m$, we associate its *Huffman reduction* to be the probability mass function $\vec{p}' = (p_1, p_2, \dots, p_{m-2}, p_{m-1} + p_m)$ over the alphabet $\{1, \dots, m-1\}$.

Fix an optimal code $C : \{1, \dots, m-1\} \to \{0,1\}^*$ for $\vec{p}'$, so that $L(C) = L_{\min}(\vec{p})$. We define an extension of C to a code

$$C^{\uparrow} : \{1, \dots, m\} \to \{0,1\}^*$$

with probability mass $\vec{p}$ by setting

$$C^{\uparrow}(i) = \begin{cases} C(i), & i < m-1 \\ C(m-1) * 0, & i = m-1, \\ C(m) * 1, & i = m, \end{cases}$$

where $*$ denotes the concatenation of words. We have

$$L(C^{\uparrow}) = \mathbb{E}_{\vec{p}}[\,\ell_{C^{\uparrow}}\,] = \mathbb{E}_{\vec{p}'}[\,\ell_C] + p_{m-1} + p_m = L_{\min}(\vec{p}) + p_{m-1} + p_m.$$

Likewise, from a canonical code $C : \{1, 2, \dots, m\} \to \{0,1\}^*$ for $\vec{p}$ we construct a code

$$C^{\downarrow} : \{1, \dots, m-1\} \to \{0,1\}^*$$

for $\vec{p}'$ by merging the codewords for the two lowest probability symbols $m-1$ and m with probabilities p_{m-1} and p_m, which are siblings by the properties of the canonical code. We obtain a single word by removing the different end bits of these siblings.

The new code $C^{\downarrow}$ has $\vec{p}'$-average length

$$L(C^{\downarrow}) = \sum_{i=1}^{m-2} p_i l_i + p_{m-1}(l_{m-1} - 1) + p_m(l_m - 1)$$
$$= L(C) - p_{m-1} - p_m = L_{\min}(\vec{p}) - p_{m-1} - p_m.$$

Suppose that C_{m-1} is an optimal code for $\vec{p}'$ and C_m is a canonical code for $\vec{p}$. We deduce from the above that

$$L(C^{\uparrow}_{m-1}) + L(C^{\downarrow}_m) = L_{\min}(\vec{p}) + L_{\min}(\vec{p}').$$

or equivalently,

$$(\,L(C^{\uparrow}_{m-1}) - L_{\min}(\vec{p})\,) = -(\,L(C^{\downarrow}_m) - L_{\min}(\vec{p}')\,)$$

The right-hand side of the above equality is non-positive since $L_{\min}(\vec{p}')$ is the optimal length for $\vec{p}'$. Likewise, the left-hand side is non-negative. Hence both sides must be equal to 0. Thus we have

$$L(C^{\downarrow}_m) = L_{\min}(\vec{p}') \;\text{ and }\; L(C^{\uparrow}_{m-1}) = L_{\min}(\vec{p}).$$

In particular, the extension of an optimal code for $\vec{p}'$ is optimal for $\vec{p}$.

As a consequence of this, if we start with an optimal code for $\vec{p}'$ with $m-1$ symbols and construct a code for m symbols by extending the codeword corresponding to $p_{m-1}+p_m$, the new code is also optimal.

We can define the Huffman codes inductively as follows.

- For an alphabet with two elements the obvious optimal code is the only Huffman code.
- Suppose we have constructed a Huffman code C_{m-1} on $\{1,\ldots,m-1\}$ with a probability mass function that is the Huffman reduction of a probability mass function on $\{1,2,\ldots,m\}$. We extend it to an optimal code $C^{\uparrow}_{m-1}$ on $\{1,\ldots,m\}$. We then transform $C^{\uparrow}_{m-1}$ to a canonical code as explained in Lemma 5.17. The resulting canonical code is a Huffman code.

The above discussion proves the following theorem, [5, p.125].

Theorem 5.19. *Let $\mathscr{X}$ be a finite set, and let p be a probability mass function on $\mathscr{X}$. Then any Huffman coding is optimal, that is, if C^* is a Huffman code and C' is any instantaneous code, $L(C^*) \le L(C')$.*

Proof. Above, we follow [5, p.25]. □

Example 5.20. Here is how the above induction works on a concrete example. Suppose that $\mathscr{X}_4 = \{1,2,3,4\}$ with associated probabilities

$$p_1 = 1/2,\ \ p_2 = 1/3,\ \ p_3 = p_4 = 1/12.$$

Its Huffman reduction is the alphabet $\mathscr{X}_3 = \{1,2,3\}$ with probabilities

$$p_1 = 1/2,\ \ p_2 = 1/3,\ \ p_3 = 1/6.$$

Finally, the Huffman reduction of $\mathscr{X}_3$ is the alphabet $\mathscr{X}_2 = \{1,2\}$ with probabilities

$$p_1 = p_2 = 1/2.$$

A Huffman code for $\mathscr{X}_2$ is $C_2(1) = 0$, $C_2(2) = 1$. We extend it to an optimal code $C_3 = C_2^{\uparrow}$ on $\mathscr{X}_3$ by setting

$$C_3(1) = 0,\ \ C_3(2) = 10,\ \ C_3(3) = 11.$$

This is already a canonical code. We extend it to a code $C_4 = C_3^{\uparrow}$ on $\mathscr{X}_4$ by setting

$$C_4(1) = C_3(1) = 0,\ \ C_4(2) = C_3(2) = 10,\ \ C_4(3) = 110,\ \ C_4(4) = 111.$$

This is already a canonical code and it is a Huffman code. The expected code length is

$$\frac{1}{2} + \frac{2}{3} + \frac{3}{12} + \frac{3}{12} = \frac{5}{3} \approx 1.666$$

The entropy of the probability mass function on $\mathscr{X}_4$ is

$$\frac{\log 2}{2} + \frac{\log 3}{3} + \frac{\log 12}{6} \approx 1.625$$

5.7. **Huffman Coding and the Twenty-Question Game.** Source coding and playing a the twenty-question game are closely related.

Consider a box containing colored balls of colors $c_1,\ldots,c_m$ with known proportions $p_1,\ldots,p_m$. A person picks a ball from the box at random. What is the most efficient sequence of yes-or-no questions to determine what colored ball the person has picked?

We claim that a sequence of questions to determine the ball is equivalent to a code for the object. A question asked in sequence will depend only on the answers to the questions asked before it in sequence, and a sequence of answers will uniquely specify the ball. Coding a yes as a 1 and a no as a 0 gives us a binary code for the set of balls, and the average code length will be the average number of answers — and therefore questions — necessary to determine a ball.

Conversely, given a binary code we can determine a sequence of yes-or-no questions corresponding to the code the answers of which will determine the ball. Note that the average number of questions here is again equal to the length of the code. A possible sequence of questions is one such that the i^{th} question asks if the i^{th} term in the ball's codeword is a 1.

Now we turn to Huffman codes. As the Huffman code is an optimal prefix code for a probability distribution and optimal sequence of questions is the sequence of questions generated by the Huffman code.

Example 5.21. Suppose we have balls of four colors $\{1, 2, 3, 4\}$ with the proportions indicated in Example 5.20

$$p_1 = 1/2, \;\; p_2 = 1/3, \;\; p_3 = p_4 = 1/12.$$

In Example 5.20 we have constructed a Huffman code

$$1 \to 0, \;\; ; 2 \to 10, \;\; 3 \to 110 \;\; 4 \to 111.$$

Here is how we ask the questions.

(i) Is the color different from 1? If the answer is 0 (or NO) we've guessed the color. It has to be red.

(ii) If the answer is 1 (or YES) it means that the ball can only be of color $2, 3$ or 4. We ask is the color different from 2. If the answer is 0 we've guessed that the ball has color 2 and we stop.

(iii) If the answer is 1 we ask is the color different from 3. If the answer is 0 we've guessed that the ball has color 3. If it is 1, the ball has color 4.

5.8. **Generation of Discrete Distributions from Fair Coins.** How many coin flips does it take to generate a random value X drawn according to some given probability mass function $\vec{p}$. First, an example.

Example 5.22. Given a sequence of fair coin tosses (which are equivalent to fair bits), suppose we wish to generate a random variable X with alphabet $\{a, b, c\}$ and probability mass distribution $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$. Given a sequence of coin flips (binary bits), we can encode a as 0, b as 10 and c as 11. Then X would have the given distribution.

We can calculate the average number of coin flips (fair bits) required to generate the distribution: $\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(2) = 1.5$ bits. This is also the entropy of the probability distribution, a result which will be investigated below.

More generally, given a sequence of fair coin tosses $F_1, F_2, \dots$ we wish to generate a discrete random variable X with range $\mathscr{X} = \{1, 2, \dots, m\}$ and probability mass function $\vec{p} = \{p_1, p_2, \dots, p_m\}$. Further, we let N denote the number of coin flips used in the algorithm.

We can use a binary tree to describe the algorithm mapping strings of bits $F_1, F_2, \dots$ to possible outcomes X. Here the leaves of the tree represent output symbols X, each node represents a 1 or a 0, and the path to the leaves gives a sequence of fair bits produced by a fair coin. This tree must satisfy certain properties:

(i) The trees should be complete. That is, every node is either a leaf or has at least two descendants in the tree. However, the tree is not necessarily finite.

(ii) The probability of a leaf at depth k is 2^{-k}. More than one leaf may be associated with the same output symbol — the total probability of all of these leaves should equal the desired probability of the output symbol.

(iii) The expected number of fair bits, $\mathbb{E}[N]$, required to generate X is equal to the expected depth of this tree.

We further note that there is more than one way to encode each random variable X - consider the code above but switching b and c. Or the example above, but with $a = 00, 11$, $b = 10$, $c = 01$. However, this last suggestion is not as efficient as the one offered in the example. This naturally gives rise to a question: what is the most efficient algorithm to generate a given distribution? Further, as we suspect, is it related to the entropy of the distribution? How so?

We expect that there would be at least as much randomness in the fair bits as we produce in the output samples. Then, as entropy is a measure of randomness and each fair bit has an entropy of 1 bit, we expect that the number of fair bits used ought to be greater than or equal to the entropy of the output — this will be addressed in the following theorem. However, we first need a lemma. Let $\mathscr{Y}$ denote the set of leaves of a complete tree and consider a distribution on the leaves such that the probability of a leaf at depth k on the tree is 2^{-k}. Let Y be a random variable drawn according to the distribution. This leads to the following lemma

Lemma 5.23. *For any complete tree, consider a probability distribution on the leaves such that the probability of a leaf at depth k is 2^{-k}. Then the expected depth of the tree is equal to the entropy of this distribution.*

Proof. We follow [5, p.136]. If we let $k(y)$ denote the depth of the leaf y then the expected depth of the tree

$$\mathbb{E}[N] = \sum_{y\in\mathscr{Y}} k(y)2^{-k},$$

and the entropy of the distribution of Y is

$$H(Y) = -\sum_{y\in\mathscr{Y}} \frac{1}{2^{k(y)}} \log \frac{1}{2^{k(y)}} = \sum_{y\in\mathscr{Y}} k(y)2^{-k}.$$

Thus we have that $\mathbb{E}[N] = H(Y)$. □

Theorem 5.24. *For any algorithm generating X, the expected number of fair bits used is greater than the entropy $H(X)$, that is*

$$\mathbb{E}[N] \geq H(X).$$

Proof. We follow [5, p.136]. Any algorithm generating X from fair bits can be represented by a complete binary tree. Label all of the leaves of this tree by $y \in \mathscr{Y} = \{1, 2, \dots\}$.

Now consider the random variable Y defined on the leaves of the tree, such that for any leaf y at depth k, the probability that $Y = y$ is 2^{-k}. By the lemma, we have

$$\mathbb{E}[N] = H(Y)$$

The random variable X is a function of Y, as one or more leaves map onto an output symbol. At this point we need to invoke the following technical result.

Lemma 5.25. *Let X be a discrete random variable taking values in a finite or countable range $\mathscr{X}$ and probability mass function $p_X(x)$. Let $g : \mathscr{X} \to \mathbb{R}$ be a function (so that $g(X)$ is a random variable taking values in $g(\mathscr{X})$ with probability mass function $p_{g(X)}(x)$). Then*

$$H(X) \geq H(g(X))$$

Proof. We follow the ideas in [5, Problem 2.4]. Observe first that the chain rule for entropy implies

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

The second equality follows by the definition of conditional entropy, Definition 2.5

$$H(g(X)|X) = \sum_x p(x)H[g(X)|X = x] = 0,$$

since $g(X)$ is determined by X.

On the other hand, the chain rule for entropy implies

$$H(g(X), X) = H(X, g(X)) = H(g(X)) + H(X|g(X)) \geq H(g(X)).$$

The inequality follows since the relative entropy is non-negative by definition. These two results combine to show the lemma. □

The above lemma shows that $H(X) \leq H(Y)$ which yields the claimed result

$$H(X) \leq \mathbb{E}[N].$$

□

The same argument answers the question of optimality for a dyadic distribution.

Theorem 5.26. *Let the random variable X have a dyadic distribution. The optimal algorithm to generate X from fair coin flips requires an expected number of coin tosses precisely equal to the entropy:*

$$\mathbb{E}[N] = H(X).$$

Proof. We follow [5, p.137]. The previous theorem shows that $H(X) \leq \mathbb{E}[N]$. We shall now show by a constructive argument that equality holds.

Consider the Huffman code tree for X as the tree to generate the random variable. For a dyadic distribution, the Huffman code is the same as the Shannon code, and achieves the entropy bound. Then for any $x \in \mathscr{X}$, the depth of the leaf in the code tree corresponding to x is the length of the corresponding codeword, which is $\log \frac{1}{p(x)}$. Thus when this code tree is used to generate X, the leaf will have probability $2^{-\log \frac{1}{p(x)}} = p(x)$. The expected number of coin flips is then the expected depth of the tree, which is equal to the entropy because the distribution is dyadic. Thus the result is achieved. □

References

[1] D. Applebaum: *Probability and Information: An Integrated Approach*, Cambridge University Press, 2008.

[2] R.B. Ash: *Information Theory* Dover Publications, 2007.

[3] R. B. Ash: *Basic Probability Theory*, Dover Publications, 2008.

[4] P. Brémaud: *Markov Chains, Gibbs Fields, Monte Carlo Simulations and Queues*

[5] T.M. Cover, M. Thomas, J. A. Thomas: *Elements of Information Theory*, Wiley-Interscience, 2006.

[6] A.J. Khinchin: *Mathematical Foundations of Information Theory* Dover, 1957.

[7] A. Renyi, Z. Makkai-Bencsath: *A Diary on Information Theory* Wiley, 1987.