Lecture Notes on Information Theory and Statistics

Shubhanshu Shekhar

Contents

Part I

The Basics

Chapter 1

Information measures for discrete distributions

In this chapter, we introduce the three main information measures (entropy, relative entropy, and mutual information) for discrete distributions with finite support. We begin this chapter with a simple guessing game, which naturally leads the definition of these three terms, and we then study some simple properties of these information measures.

Notation. As mentioned earlier, throughout this chapter, we will work with discrete distributions supported on a finite set, which we denote by $\mathcal{X}$. The elements of this set could be anything, and in particular, we do not assume that there exists any ordering among them. We will mostly consider log to the base 2, unless otherwise stated.

1.1 A Guessing Game

We consider a collection of simple guessing games (or a 20 questions games) that will motivate the definitions of the three main information measures: entropy, relative entropy, and mutual information.

Question 1 (Guessing Game I)**.** *In the simplest setting, suppose there are n identical bins, and a ball is placed uniformly at random in one of the bins. Denote the position of the ball with the random variable $X \sim Uniform(\mathcal{X})$, where $\mathcal{X} = \{0, 1, \ldots, n-1\}$. Suppose, we can make binary queries of the form:* "Is X in the set A?" *for $A \subset \mathcal{X}$. Our objective is to design a strategy of asking a series of such questions, such that the average number of questions required to identify the bin containing the ball (i.e., the value of the random variable X) is small.*

A simple strategy could be to ask the series of questions: is $X = 0$, is $X = 1$, and so on. With this strategy, we can check that the average number of questions needed are equal to

$$L = \sum_{i=0}^{n-1} \mathbb{P}(X = i)(i+1) = \frac{1}{n}\sum_{i=0}^{n-1}(i+1) = \frac{1}{n}\left(1 + 2 + \ldots + n\right) = \frac{n}{2}.$$

Thus, the number of yes/no questions required by this strategy (called the linear search) is $\Omega(n)$. This strategy is quite inefficient, since with every query we reduce the search space by one. As we see next, we can do significantly better.

An optimal strategy for the above problem is the *binary search*, in which the queries are specifically designed to reduce the size of search space by half with each query. In particular, consider the case of $n = 4$. Then, the binary search decision tree is shown in Figure 1.1. Assigning the values $0 \leftarrow Y$ and $1 \leftarrow N$, we see that the series of questions to ascertain a value of $X = i$ is equivalent to the binary encoding or representation of i.

$$0 \equiv (YY) \equiv (00), \quad 1 \equiv (YN) \equiv (01), \quad 2 \equiv (NY) \equiv (10), \quad 3 \equiv (NN) \equiv (11).$$

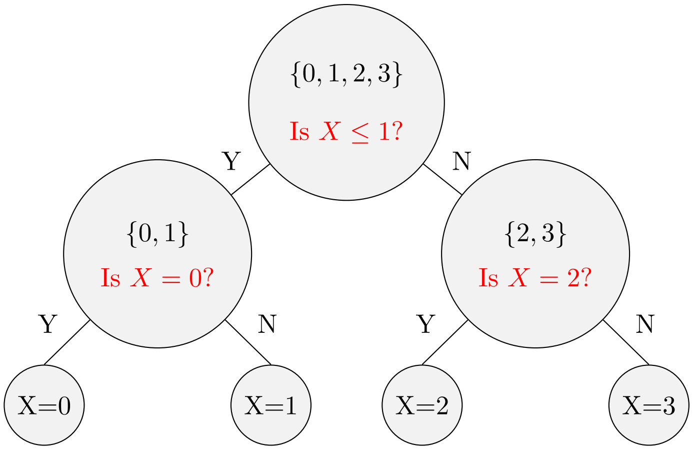

Figure 1.1: The figure shows the decision tree for the optimal strategy (i.e., binary search) for the guessing game with X drawn uniformly at random from $\mathcal{X} = \{0,1,2,3\}$. Each query is chosen to ensures that the outcomes (i.e., Y or N) are equiprobable. In other words, this strategy proceeds by greedily selecting a query whose outcome is the most uncertain.

Denote the number of questions needed to verify $X = i$ by ℓ_i. Then, the binary search scheme asks $\ell_i = 2$ questions for all $i \in \mathcal{X}$ (in other words, it represents each $i \in \mathcal{X}$ with a *codeword* of length $\ell_i = 2$). Interestingly, 2 is also equal to the negative of the logarithm of the probability assigned to each value $i \in \mathcal{X}$ to the base 2; that is, $2 = \log(1/p_i) = \log(4)$. The average number of questions required by this (optimal) strategy is then equal to

$$L = 2 = \sum_{i=0}^{3} p_i \ell_i = \sum_{i=0}^{3} p_i \log(1/p_i).$$

Thus, the above discussion suggests that the number of binary questions needed to completely remove the uncertainty about the value of $X \sim Uniform(\mathcal{X})$ is $\log(|\mathcal{X}|)$. What is the analog of this quantity for non-uniform distribution over $\mathcal{X}$? We consider this question in the next version of the guessing game.

Question 2 (Guessing game II)**.** *Consider the same setting as Question 1 with $\mathcal{X} = \{0,1,2,3\}$, but assume that X is drawn from the following distribution (instead of uniformly):*

$$P_X = (p_0, p_1, p_2, p_3) = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}\right).$$

What is the optimal sequence of binary questions to identify the true value of X?

We will develop a strategy for this problem, motivated by the 'halving' property of the binary search scheme for the uniform case. In particular, we will take the strategy which can be summarized as follows:

make queries whose outcomes are equally likely (or in other words, are most uncertain).

Note that when X is uniformly distributed, the above strategy reduces exactly to the binary search. The decision tree of this strategy for the distribution of Question 2 is shown in Section 1.1. Unlike the previous game, the decision tree is not balanced — it asks more questions of the less likely values of i. The expected number of questions in this case is equal to

$$\begin{aligned} L &= \sum_{i=0}^{3} p_i \ell_i = \frac{1}{8} \times 3 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{2} \times 1 = \frac{15}{8} \\ &= -\sum_{i=1}^{n} p_i \log p_i := H(P_X). \end{aligned}$$

Again the average number of yes/no questions needed to learn X is characterized by the quantity, $-\sum_{i\in\mathcal{X}} p_i \log p_i$. This functional of the probability distribution is called its *entropy*, also called its self-information. As we will see later, it is a fundamental limit on the average number of yes/no questions needed to learn the value of X (or equivalently, the average length of a binary lossless representation of all realizations of X).

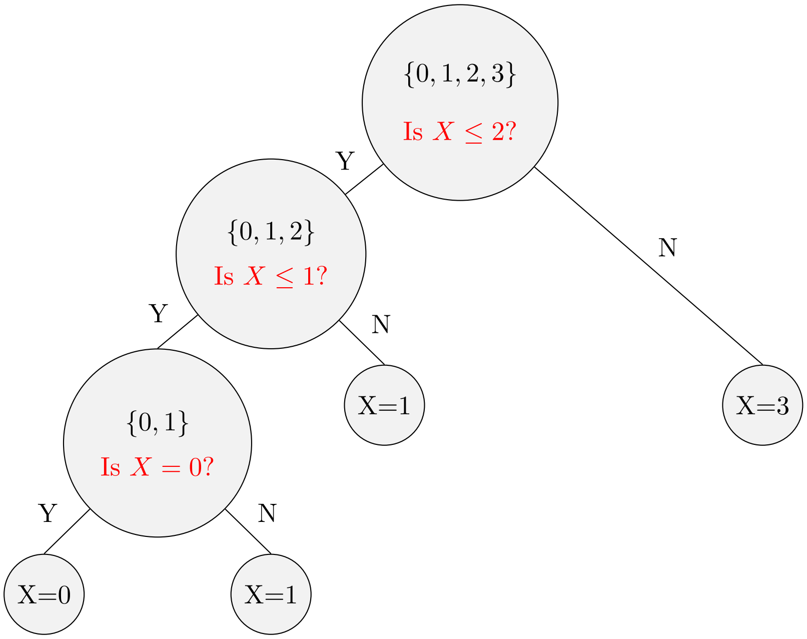

Figure 1.2: The figure shows the decision tree for the optimal strategy for a non-uniform probability distribution over $\mathcal{X}$. The key observation is that this strategy assigns fewer questions (or a shorter codeword) to the higher probability symbol (3), and more questions to the less probable symbols, such as 0 and 1.

In both the previous games, we assumed that the player knows the true distribution of X exactly. We now consider a situation where there is a mismatch between the true and the assumed distribution of X.

Question 3. *Consider the same setting as Question 2, but now assume that the true distribution (P_X) of X is not known to the player, and instead he believes that $X \sim Q_X$, where $Q_X = (1/2, 1/8, 1/8, 1/4)$. What is the effect of this distribution mismatch?*

Under the assumption that the true distribution is Q_X, the optimal codeword assignment is

$$0 \equiv (Y) \equiv (0), \quad 1 \equiv (NNY) \equiv (110), \quad 2 \equiv (NNN) \equiv (111), \quad 3 \equiv (NY) \equiv (10).$$

Again, as before, in this example, the number of questions needed to ascertain that $X = i$ is equal to $\log(1/q_i)$. The average number of questions needed to learn the value of X is

$$\begin{aligned} L = \sum_{i=0}^{3} p_i \log(1/q_i) &= -\sum_{i=1}^{3} p_i \log(p_i) + \sum_{i=0}^{3} p_i \log(p_i/q_i) \\ &= H(P_X) + D_{\mathrm{kl}}(P_X \parallel Q_X). \end{aligned}$$

The second term in the display is called the relative entropy or KL divergence between P_X and Q_X, and it denotes the price paid by the player for using the wrong model for asking the yes/no questions (or the extra average codeword length incurred due to the ignorance of the true distribution). As we will see later, this quantity is always non-negative.

Question 4. *Finally, we now consider a case of guessing another random variable Y on the set $\{a, b\}$. The joint distribution of X and Y is stated in Table 1.1.*

Suppose we want to ask a series of yes/no questions to find out the true value of both X and Y. Consider two strategies:

	$X=0$	$X=1$	$X=2$	$X=3$
$Y=a$	0	1/8	1/8	0
$Y=b$	1/8	0	1/8	1/2

Table 1.1: Joint distribution P_{XY} of (X,Y).

- *Player 1 develops two independent strategies for learning X and Y*
- *Player 2 develops a joint strategy for learning X and Y together.*

Who does better in terms of the average number of questions needed to learning both X and Y? How much is the improvement?

From the table, we can check that Y has a distribution $P_Y = (1/4, 3/4)$ over the set $\{a, b\}$. The marginal distribution of X is the same as in the previous two questions. Hence, we expect that the average number of yes/no questions need by player 1 (denoted by L_1) is

$$L_1 = H(P_X) + H(P_Y) \approx 1.75 + 0.811 \text{ bits} = 2.561 \text{ bits}$$

For the second player, who develops a joint strategy for querying about (X,Y), we expect the average number of yes/no questions (denoted by L_2) to be

$$L_2 = H(P_X, P_Y) = 4 \times \frac{1}{8} \times 3 + \frac{1}{2} \times 1 = 2 \text{ bits}.$$

Thus, the player who jointly considers the two random variables requires fewer questions. This is because, knowing the value of Y also provides information about the X value. For instance, if we know that $Y = a$, then we know that X cannot be 0 or 3. Exploiting this leads to the reduced number of questions needed by player 2. The amount of improvement, $L_1 - L_2$, is equal to

$$L_1 - L_2 = H(X) + H(Y) - H(X,Y) := I(X;Y).$$

The term $I(X;Y)$ is called the *mutual information* between the random variables X and Y, and it precisely quantifies the amount of information that X contains about Y (or equivalently, Y contains about X; since $I(X;Y)$ is symmetric).

Summary. The above discussion can be summarized as follows:

- Entropy $H(X) = -\sum_i p_i \log(p_i)$ quantifies the information content of a random variable X. It is also equal to the minimum average number of yes/no questions needed to learn the value of X.
- The relative entropy is a measure of discrepancy between two distributions P_X and Q_X. It quantifies the additional (on an average) number of yes/no questions needed to learn about X, under wrong model assumptions.
- The mutual information $I(X;Y)$ is measure of dependence between X and Y. It quantifies how much information about X is contained in the random variable Y.

1.2 Formal Definitions

In this section, we formally define the three information measures, and observe some of their basic properties. The next section contains a more thorough treatment of the properties of these measures.

Definition 5. Suppose X denotes an $\mathcal{X}$-valued random variable with probability mass function (p.m.f.) p_X. Then, the entropy of X (actually the distribution p_X) is defined as

$$H(X) \equiv H(p_X) = \sum_{x \in \mathcal{X}} -p_X(x) \log(p_X(x)). \tag{1.1}$$

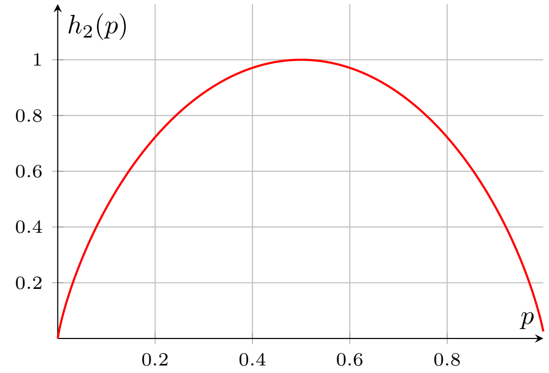

Figure 1.3: Variation of the binary entropy $h_2(p) = -p\log(p) - \bar{p}\log(\bar{p})$ with the parameter p. Note that $h_2(p)$ is zero for $p \in \{0,1\}$, when the random variable is deterministic, and is maximized at $p = 0.5$, which corresponds to the maximum uncertainty. Also note that the qualitative behavior of the curve is similar to that of the variance $p(1-p)$; another measure of dispersion or variability of the distribution.

For a pair of distributions (X, Y) on $\mathcal{X} \times \mathcal{X}$ with joint distribution p_{XY}, their joint entropy is defined as, following (1.1),

$$H(X,Y) \equiv H(p_{XY}) = \sum_{x\in\mathcal{X}} \sup_{y\in\mathcal{X}} -p_{XY}(x,y)\log(p_{XY}(x,y)).$$

Finally, the conditional entropy of X given Y is defined as the average (over the marginal p_X) of the entropy of $Y|X = x$. That is,

$$H(Y|X) \equiv H(p_{Y|X}|p_X) = \sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{X}} -p_{Y|X}(y|x)\log\left(p_{Y|X}(y|x)\right).$$

Remark 6. The base of the logarithm used in defining the entropy characterizes its unit: if the base is 2, entropy is measured in *bits*, while for natural logarithm with base e, the unit is called *nats*. It is easy to verify that changing the base from a to b is achieved by the simple identity: $H_a(X) = \log_b a H_b(X)$.

Entropy is defined as a quantitative measure of the amount of 'uncertainty' contained in a probability distribution. In fact, there exist several results that begin by stating a set of reasonable axioms to be satisfied by a good measure of uncertainty, and then prove that the above definition is the only one that simultaneously satisfies those properties [Csiszár and Körner, 2011].

We can also intuitively see that the above definition serves as a good measure of uncertainty. For instance, suppose $\mathcal{X} = \{0,1\}$ and X is a Bernoulli distribution with parameter p. Then, the entropy of X, also called binary entropy, and denoted by $h_2(p)$, is defined as

$$h_2(p) = -p\log(p) - \bar{p}\log\bar{p}, \quad \text{where } \bar{p} := 1-p.$$

On plotting it, we can see that the $h_2(p)$ is zero at $p \in \{0,1\}$, and it achieves its maximum at $p = 1/2$.

The definition of entropy leads to some immediate conclusions that we record next:

Proposition 7. *The following statements are true for $\mathcal{X}$-valued random variables X, Y etc.:*

(a) $H(X) \geq 0$, for all random variables X, and its minimum value of 0 is achieved if and only if X is equal to a constant with probability 1.

(b) The joint entropy of (X,Y) is equal to the entropy of X, plus the conditional entropy of Y given X:

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

(c) For any function $f : \mathcal{X} \to \mathcal{X}$, we have $H(f(X)) \leq H(X)$, with equality iff f is bijective.

Proof. (a) Since $H(X) = \sum_{x \in \mathcal{X}} p(x) \log 1/p(x)$, it is a sum of non-negative terms, which implies that $H(X) \geq 0$. To achieve the equality, note that each $p(x) \log 1/p(x)$ must be equal to zero; which implies that for all $x \in \mathcal{X}$, the value of $p(x)$ must lie in $\{0, 1\}$. Since $\sum_{x \in \mathcal{X}} p(x)$ is constrained to be equal to 1, the result follows.

(b) This follows directly by the definition of joint and conditional entropies. In particular,

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

Repeating the same argument, but now writing $p(x,y) = p(y)p(x|y)$, we get the other equivalent definition.

(c) This is a simple consequence of the previous two results. In particular, we set $Y = f(X)$. Then,

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

where the last equality is a consequence of the fact that $Y = f(X)$; and hence $H(Y|X)$ is equal to 0. Since $H(X|Y)$ is not necessarily zero, we get the required inequality $H(Y) \leq H(X)$. In words, this means that we cannot add more uncertainty (or self-information) to a signal by applying a deterministic function.

□

We now present a simple application of the law of large numbers to characterize the support of a high dimensional product distribution in terms of the entropy.

Proposition 8 (Asymptotic equipartition probability (AEP). *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_X$, and for a fixed $\epsilon > 0$, define the 'typical set' $A_n(\epsilon)$ as*

$$A_n(\epsilon) = \left\{ x^n \in \mathcal{X}^n : 2^{-n(H(X)-\epsilon)} \leq \mathbb{P}(X^n) = \prod_{i=1}^n P_X(X_i) \leq 2^{-n(H(X)+\epsilon)} \right\}.$$

Then, the following statements are true:

- $\lim_{n \to \infty} \mathbb{P}(A_n(\epsilon)) = 0$.
- $|A_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$.
- *For n large enough, we have $|A_n(\epsilon)| \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$.*

Thus, for large n, a small subset $A_n(\epsilon)$ of $\mathcal{X}^n$, consisting of roughly equiprobable sequences, contains almost all the probability.

Proof.
- This is simply an application of the weak LLN to $-\log(\mathbb{P}(X^n))$.
- We prove this by noting that $\mathbb{P}(A_n(\epsilon)) \leq 1$, and thus

$$1 \geq \sum_{x^n \in A_n(\epsilon)} \mathbb{P}(x^n) \geq \sum_{x^n \in A_n(\epsilon)} 2^{-nH(X)-n\epsilon} = |A_n(\epsilon)| 2^{-nH(X)-n\epsilon}.$$

- By the definition of convergence in probability, for n large enough, we have $\mathbb{P}(A_n(\epsilon)) \geq 1 - \epsilon$. Hence, we have

$$1 - \epsilon \leq \mathbb{P}\left(A_n(\epsilon)\right) = \sum_{x^n \in A_n(\epsilon)} \mathbb{P}(x^n) \leq \sum_{x^n \in A_n(\epsilon)} 2^{-n(H(X)-\epsilon)} = |A_n(\epsilon)| 2^{-n(H(X)-\epsilon)}.$$

□

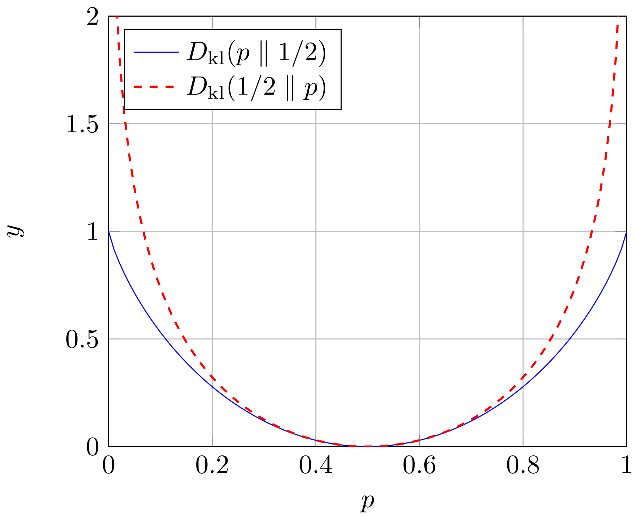

Figure 1.4: Plots of $D_{\mathrm{kl}}(P_X \parallel P_Y)$ and $D_{\mathrm{kl}}(P_Y \parallel P_X)$, where $X \sim \text{Bernoulli}(p)$ and $X \sim \text{Bernoulli}(1/2)$, as p varies in $[0, 1]$. The figures illustrates the locally quadratic behavior of relative entropy, and also shows that it is not symmetric.

Remark 9. The above result can be used to construct a theoretically simple, but computationally infeasible, compression scheme. The idea is simple: enumerate all elements of $x^n \in A_n(\epsilon)$, and assign them their binary representation prefixed by an additional '0' as the codeword; and for all points in $A_n(\epsilon)^c$, assign a codeword starting with '1', followed by the binary representation after enumerating all elements. It is easy to check that this lossless compression scheme has an average codeword length (per symbol) smaller than $H(X) + (2 + \epsilon + \log(|\mathcal{X}|))/n$.

The next, and perhaps the most important, information measure that we introduce is the relative entropy, also known as the Kullback-Leibler or KL divergence.

Definition 10 (Relative Entropy)**.** The relative entropy between two distributions P_X and Q_Y on the same domain $\mathcal{X}$ is defined as

$$D_{\mathrm{kl}}(P_X \parallel Q_Y) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

For joint distributions P_{XY} and Q_{XY}, the conditional relative entropy is defined as

$$D_{\mathrm{kl}}\left(P_{Y|X} \parallel Q_{Y|X} | P_X\right) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \left(\frac{p(y|x)}{q(y|x)} \right) = \sum_{x \in \mathcal{X}} p(x) D_{\mathrm{kl}}\left(P_{Y|X=x} \parallel Q_{Y|X=x}\right).$$

In other words, the conditional relative entropy is the average (over the marginal P_X) relative entropy between the conditional distributions $P_{Y|X=x}$ and $Q_{Y|X=x}$.

From the definition we can see that the relative entropy is not symmetric in its arguments. As an example, consider the two ways of computing the relative entropy between $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{Bernoulli}(1/2)$, as plotted in Figure 1.4.

Furthermore, if $P \not\ll Q$, then $D_{\mathrm{kl}}(P, Q) = \infty$. We note some other basic properties of relative entropy in our next result.

Proposition 11. *(a) Let $\varphi : [0, \infty) \to \mathbb{R}$ denote the function $\varphi(x) = x \log x$. Then, for two distributions P and Q with p.m.f. p and q respectively, we have $D_{kl}(P \parallel Q) = \mathbb{E}_Q[\varphi(p(X)/q(X))]$.*

(b) For any two distributions P and Q, we have $D_{kl}(P \parallel Q) \geq 0$. The equality holds if and only if $P = Q$.

(c) Let U denote the uniform distribution over (the finite set) $\mathcal{X}$. Then, we have

$$D_{kl}(P \parallel U) = \log(|\mathcal{X}|) - H(P).$$

Proof. (a) Let $\ell(x)$ denote $p(x)/q(x)$. Then, the relative entropy between P and Q can be written as

$$\begin{aligned} D_{\mathrm{kl}}(P \parallel Q) &= \sum_{x\in\mathcal{X}} p(x) \log\left(\ell(x)\right) = \sum_{x\ in\mathcal{X}} q(x)\ell(x)\log\ell(x) \\ &= \sum_{x\in\mathcal{X}} q(x)\varphi(x) = \mathbb{E}_Q[\varphi(\ell(X))]. \end{aligned}$$

(b) It is easy to verify that the mapping $x \mapsto \varphi(x)$ is convex. Hence, by an application of Jensen's inequality, we have

$$\begin{aligned} D_{\mathrm{kl}}(P \parallel Q) = \mathbb{E}_Q[\varphi(\ell(X))] \geq \varphi\left(\mathbb{E}_Q[\ell(X)]\right) = \varphi\left(\sum_{x\in\mathcal{X}} q(x)\frac{p(x)}{q(x)}\right) \\ = \varphi(1) = 0. \end{aligned}$$

The above inequality holds with equality if and only if the function $\ell(x)$ is a constant. In other words, this is an equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

(c) This result follows directly by definition. In particular, we have

$$\begin{aligned} D_{\mathrm{kl}}(P \parallel U) &= \sum_{x\in\mathcal{X}} p(x)\log\big(|\mathcal{X}|p(x)\big) = \log(|\mathcal{X}|)\sum_{x\in\mathcal{X}} p(x) + \sum_{x\in\mathcal{X}} p(x)\log p(x) \\ &= \log(|\mathcal{X}|) - H(P). \end{aligned}$$

□

An immediate corollary of the above result is that the distribution over $\mathcal{X}$ that achieves the maximum entropy is the uniform distribution. Our next result shows that this statement is more generally true: operations that make a distribution even *slightly more uniform* lead to an increase in entropy.

Proposition 12. *Given a distribution p on $\mathcal{X}$, let q denote a distribution which replaces $p(x)$ and $p(x')$ with $0.5(p(x)+p(x'))$. Then, we have $H(q) \geq H(p)$.*

Proof. The proof of this result is a direct consequence of Jensen's inequality. In particular, note that due to the convexity of the function $\varphi(x) = x\log x$, we have

$$\begin{aligned} q(x)\log q(x) &= \frac{p(x)+p(x')}{2}\log\left(\frac{p(x)+p(x')}{2}\right) \leq \frac{1}{2}\left(p(x)\log p(x) + p(x')\log p(x')\right) \\ q(x')\log q(x') &= \frac{p(x)+p(x')}{2}\log\left(\frac{p(x)+p(x')}{2}\right) \leq \frac{1}{2}\left(p(x)\log p(x) + p(x')\log p(x')\right). \end{aligned}$$

On adding the two inequalities, we get

$$-q(x)\log q(x) - q(x')\log q(x') \geq -p(x)\log p(x) - p(x')\log p(x').$$

Since the two entropies differ only on the terms x and x', the result follows. □

Finally, we introduce the third important quantity, the mutual information.

Definition 13 (Mutual Information)**.** The mutual information between two random variables, X and Y on the domain $\mathcal{X}$, is defined as

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

Alternatively, it is also defined as the relative entropy between the joint distribution of (X, Y) and the product of their marginals:

$$I(X;Y) = D\left(P_{XY} \parallel P_X \times P_Y\right).$$

The conditional mutual information between X and Y given Z, is defined as

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

It is easy to verify that the two definitions are equivalent. Interestingly, the second definition immediately implies that unlike relative entropy, the mutual information is symmetric quantity. The first definition provides an intuitive explanation of mutual information: since entropy is a measure of uncertainty, the definition suggests that the mutual information between X and Y is the *reduction in uncertainty* about X that is achieved (on average) with the knowledge of Y.

Proposition 14. *(a) All the definitions of mutual information in Definition 13 are equivalent.*

(b) For any random variable X, we have $I(X;X) = H(X)$. Hence, entropy is also known as the self information.

(c) $I(X;Y) \geq 0$, and $I(X;Y) = 0$ if and only if $X \perp Y$.

(d) Conditioning reduces entropy: for any pair (X, Y), we have $H(X|Y) \leq H(X)$, with equality if and only if $X \perp Y$.

Proof. (a) The first two definitions follow from the definition of joint entropy. In particular, since $H(X, Y) = H(X) + H(Y|X)$, we have

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

Similarly, using $H(X,Y) = H(Y) + H(X|Y)$, we get the other definition $I(X;Y) = H(X) - H(X|Y)$.

To get the relative entropy definition, we start with $I(X;Y) = H(X) + H(Y) - H(X,Y)$, to get

$$\begin{aligned} I(X;Y) &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{y\in\mathcal{Y}} p(y)\log p(y) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= D_{\mathrm{kl}}\left(P_{XY} \parallel P_X P_Y\right). \end{aligned}$$

(b) This follows directly from the definition: $I(X;X) = H(X) - H(X|X) = H(X)$.

(c) The nonnegativity follows from the relative entropy definition. Furthermore, equality occurs if and only if $P_{XY} = P_X \times P_Y$, which is equivalent to independence of X and Y.

(d) This is a direct consequence of the nonnegativity: $0 \leq I(X;Y) = H(X) - H(X|Y)$, which implies that $H(X|Y) \leq H(X)$ as needed. Furthermore, since $I(X;Y)$ is equal to zero iff $X \perp Y$, we have $H(X|Y) = H(X)$ iff $X \perp Y$.

□

The above results imply that mutual information $I(X;Y)$ serves as a measure of dependence between the two random variables, and its value ranges from 0 (for $X \perp Y$) to $H(X)$ (when $X = Y$).

1.2.1 An application: generating purely random bits

Even with the elementary properties we have seen so far, we can obtain non-trivial results about important practical applications, such as the task of generating purely random bits (i.e., fair coin tosses) from bent coins.

Question 15. *Suppose $X \sim$ Bernoulli(p) with $p \in [0,1]$ unknown: that is, X denotes the outcome of a* bent *coin. Can we use this bent coin to obtain a fair coin toss $Y \sim$ Bernoulli$(1/2)$?*

Here is a simple procedure, that can simulate a fair coin toss using the bent coin:

- Toss the bent coin twice, and let (X_1, X_2) denote the outcomes. Define the event $E = \{(X_1, X_2) = (0,1)$ or $(X_1, X_2) = (1,0)\}$.
- On observing the outcomes, if event E occurs, then define $H \equiv \{(X_1, X_2) = (1,0)\}$ and $T \equiv \{(X_1, X_2) = (0,1)\}$. Then, we have

  $$P(H|E) = P(H \cap E)/P(E) = \frac{\mathbb{P}\left(X_1 = 1, X_2 = 0\right)}{\mathbb{P}\left(X_1 = 1, X_2 = 0\right) + \mathbb{P}\left(X_1 = 0, X_2 = 1\right)} = \frac{p(1-p)}{p(1-p) + (1-p)p} = \frac{1}{2}.$$

  Similarly, we can show that $P(T|E) = 1/2$.
- If the event E does not occur, then repeat the process.

Question 16. *How can we generalize this scheme to extract multiple uniform bits from n i.i.d. draws from the bent coin? What is the average number of random bits we can extract from n i.i.d. draws?*

Given n draws of the bent coin, denote by $X_1, \ldots, X_n$ (and assuming that n is event), we proceed as follows:

- Set $K = 0$.
- For i in the range $\{1, 2, \ldots, n/2\}$, observe the pair (X_{2i-1}, X_{2i}), and do one of two things:
  - If $(X_{2i-1}, X_{2i}) \in \{(1,0), (0,1)\}$, then set Z_{K+1} equal to 1 or 0 according to the previous rule. Increment $K \leftarrow K + 1$.
  - Otherwise, discard the pair (X_{2i-1}, X_{2i}).
- Return the fair coin tosses $(Z_1, Z_2, \ldots, Z_K)$.

It is easy to check that for this scheme, conditioned on $K = k$, all the 2^k bits are equally likely. Furthermore, the expected number of purely random bits that we can extract from n bent coin tosses is

$$\mathbb{E}[K] = \sum_{i=1}^{n/2} \mathbb{P}\left((X_{2i-1}, X_i) \in \{(1,0), (0,1)\}\right) = np(1-p).$$

Question 17. *Can we do better?*

We can obtain a method agnostic upper bound on the achievable performance of any random bit generating scheme $\mathcal{A}$. In particular, let $\mathcal{A}$ denote any scheme that maps $X^n = (X_1, \ldots, X_n)$ to a random bits $(Z_1, Z_2, \ldots, Z_K, K)$, which satisfy the property: *conditioned on $K = k$, the vector Z^k is uniformly distributed over $\{0,1\}^k$, for all $k \in [n]$.* We then have the following:

$$\begin{aligned} nh_2(p) = H(X^n) &\geq H(Z^K, K) = H(K) + H(Z^K|K) \\ &\geq H(Z^K|K) = \sum_{k \geq 1} \mathbb{P}(K = k) H(Z^k|K = k) \\ &= \sum_{k \geq 1} \mathbb{P}(K = k) k = \mathbb{E}[K]. \end{aligned}$$

There exist methods, such as Peres' iterated extractor [Peres, 1992], that achieves this optimal rate in the limit of large n.

1.3 Properties of Information Measures

1.3.1 Chain Rules

We begin by establishing the chain rules for the three information measures. These results decompose the information measures for a collection of random variables into a sum of simpler terms.

Theorem 18 (Chain rule for entropy)**.** *Consider a random vector* $(X_1, X_2, \ldots, X_n)$ *taking values in* $\mathcal{X}^n$*. Then, we have*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots H(X_i|X_1, \ldots, X_{i-1}) + \ldots + H(X_n|X_1, \ldots, X_{n-1}).$$

For any permutation $\pi : [n] \to [n]$*, we have*

$$H(X_1, \ldots, X_n) = H(X_{\pi(1)}) + H(X_{\pi(2)}|X_{\pi(1)}) + \ldots + H(X_{\pi(n)}|X_{\pi(1)}, \ldots, X_{\pi(n-1)}).$$

Since conditioning reduces entropy, we also have the following inequality:

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

with equality only if $(X_1, \ldots, X_n)$ *are independent.*

Proof. The proof follows by induction:

- For $n = 2$, we already proved it directly from the definition of joint entropy.
- Suppose the statement is true for $n - 1$.
- Then, we have

$$\begin{aligned} H(X_1, \ldots, X_n) = H(X^{n-1}, X_n) &\overset{(i)}{=} H(X^{n-1}) + H(X_n|X^{n-1}) \\ &\overset{(ii)}{=} \sum_{i=1}^{n-1} H(X_i|X^{i-1}) + H(X_n|X^{n-1}) = \sum_{i=1}^{n} H(X_i|X^{n-1}), \end{aligned}$$

where (i) follows from the $n = 2$ case, and (ii) follows from the induction hypothesis.

□

Theorem 19 (Chain rule for relative entropy.)**.** *Consider two joint distribution of* $(X_1, \ldots, X_n)$*, denoted by* P_{X^n}*, and* Q_{X^n}*. Then, the relative entropy between these two distributions can be decomposed as*

$$D_{kl}(P_{X^n} \parallel Q_{X^n}) = \sum_{i=1}^{n} D_{kl}\left(P_{X_i|X^{i-1}} \parallel Q_{X_i|X^{i-1}}|P_{X^{i-1}}\right). \tag{1.2}$$

Again, for any permutation $\pi : [n] \to [n]$*, we have*

$$D_{kl}(P_{X^n} \parallel Q_{X^n}) = \sum_{i=1}^{n} D_{kl}\left(P_{X_{\pi(i)}|X_{\pi(1)}^{\pi(i-1)}} \parallel Q_{X_{\pi(i)}|X_{\pi(1)}^{\pi(i-1)}}|P_{X_{\pi(1)}^{\pi(i-1)}}\right).$$

If $Q_{X^n} = \prod_{i=1}^{n} Q_{X_i}$*, then, we have*

$$\begin{aligned} D_{kl}\left(P_{X^n} \parallel Q_{X^n}\right) &= D_{kl}\left(P_{X^n} \parallel \prod_{i=1}^{n} P_{X_i}\right) + \sum_{i=1}^{n} D_{kl}\left(P_{X_i} \parallel Q_{X_i}\right) && (1.3) \\ &\overset{(a)}{\geq} \sum_{i=1}^{n} D_{kl}\left(P_{X_i} \parallel Q_{X_i}\right). && (1.4) \end{aligned}$$

The equality in (a) *occurs when* P_{X^n} *is equal to the product of its marginals.*

Proof. We prove (1.2) for the special case $n = 2$, since the general case follows by induction.

$$\begin{aligned}
D_{\mathrm{kl}}(P_{X_1X_2} \parallel Q_{X_1X_2}) &= \sum_{x_1,x_2} p(x_1,x_2)\log\left(\frac{p(x_1,x_2)}{q(x_1,x_2)}\right) = \sum_{x_1,x_2} p(x_1,x_2)\log\left(\frac{p(x_1)p(x_2|x_1)}{q(x_1)q(x_2|x_1)}\right)\\
&= \sum_{x_1} p(x_1)\log\left(\frac{p(x_1)}{q(x_1)}\right) + \sum_{x_1,x_2} p(x_1,x_2)\log\left(\frac{p(x_2|x_1)}{q(x_2|x_1)}\right)\\
&= D_{\mathrm{kl}}(P_{X_1} \parallel Q_{X_1}) + + \sum_{x_1} p(x_1)\sum_{x_2} p(x_2|x_1)\log\left(\frac{p(x_2|x_1)}{q(x_2|x_1)}\right)\\
&= D_{\mathrm{kl}}(P_{X_1} \parallel Q_{X_1}) + D_{\mathrm{kl}}(P_{X_2|X_1} \parallel Q_{X_2|X_1}|P_{X_1}).
\end{aligned}$$

Next, to prove (1.3), we note that

$$\begin{aligned}
D_{\mathrm{kl}}(P_{X^n} \parallel Q_{X^n}) &= \sum_{x^n} p(x^n)\log\left(\frac{p(x^n)}{\prod_{i=1}^n q_i(x_i)}\right) = \sum_{x^n} p(x^n)\log\left(\frac{p(x^n)}{\prod_{i=1}^n p_i(x_i)}\frac{\prod_{i=1}^n p_i(x_i)}{\prod_{i=1}^n q_i(x_i)}\right)\\
&= \sum_{x^n} p(x^n)\log\left(\frac{p(x^n)}{\prod_{i=1}^n p_i(x_i)}\right) + \sum_{x^n} p(x^n)\log\left(\frac{\prod_{i=1}^n p_i(x_i)}{\prod_{i=1}^n q_i(x_i)}\right)\\
&= D_{\mathrm{kl}}(P_{X^n} \parallel \prod_{i=1}^n P_{X_i}) + \sum_{i=1}^n \sum_{x_i} p_i(x_i)\log\left(\frac{p_i(x_i)}{q_i(x_i)}\right)\\
&= D_{\mathrm{kl}}(P_{X^n} \parallel \prod_{i=1}^n P_{X_i}) + \sum_{i=1}^n D_{\mathrm{kl}}(P_{X_i} \parallel Q_{X_i}).
\end{aligned}$$

The inequality (1.4) follows due to the nonnegativity of $D_{\mathrm{kl}}(P_{X^n} \parallel \prod_{i=1}^n P_{X_i})$, and note that the equality occurs if and only if this term is zero. This happens iff $P_{X^n} = \prod_{i=1}^n P_{X_i}$. □

Corollary 20. *A simple consequence of the chain rule for relative entropy is the fact that "conditioning increases divergence". Namely, suppose $P_{XY} = P_X P_{Y|X}$, and $Q_{XY} = P_X Q_{Y|X}$. Then, we have*

$$\begin{aligned}
D_{kl}\left(P_{XY} \parallel Q_{XY}\right) &= D_{kl}(P_Y \parallel Q_Y) + D_{kl}\left(P_{X|Y} \parallel Q_{X|Y}|P_Y\right)\\
&= D_{kl}\left(P_X \parallel P_X\right) + D_{kl}\left(P_{Y|X} \parallel Q_{Y|X}|P_X\right).
\end{aligned}$$

Since $D_{kl}(P_X \parallel P_X) = 0$ and $D_{kl}(P_{X|Y} \parallel Q_{X|Y}|P_Y) \geq 0$, the above implies

$$D_{kl}(P_Y \parallel Q_Y) \leq D_{kl}\left(P_{Y|X} \parallel Q_{Y|X}|P_X\right).$$

Finally, since mutual information is an instance of relative entropy, it also satisfies an analogous chain rule.

Proposition 21 (Chain rule for mutual information)**.**

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

For any permutation $\pi : [n] \to [n]$, we have

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

Proof. This is a simple consequence of the chain rule for entropy. Again, we prove the result for the case of $n = 2$, since the general result follows by induction.

$$\begin{aligned}
I(X_1,X_2;Y) &= H(X_1,X_2) - H(X_1,X_2|Y) = H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|X_1,Y)\\
&= \big(H(X_1) - H(X_1|Y)\big) + \big(H(X_2|X_1) - H(X_2|X_1,Y)\big)\\
&= I(X_1;Y) + I(X_2;Y|X_1).
\end{aligned}$$

□

Application: Han's inequality and an isoperimetric inequality for the binary hypercube. The simple chain rules obtained in this section often serve as an important tool in establishing various properties of information measures. We illustrate this by proving Han's inequalities for entropy and relative entropy.

Proposition 22. *Let $X_1, \ldots, X_n$ denote n, possibly dependent, $\mathcal{X}$-valued random variables. For any $i \in [n]$, define $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Then, we have*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X^{(i)}).$$

Proof. Introduce the notations $X^i = (X_1, \ldots, X_i)$ and $X_i^j = (X_i, \ldots, X_j)$ for $j \geq i$. Then, we have

$$\begin{aligned} H(X^n) &\overset{(i)}{=} H(X^{(i)}) + H(X_i|X^{(i)}) = H(X^{(i)}) + H(X_i|X^{i-1}, X_{i+1}^n) \\ &\overset{(i)}{\leq} H(X^{(i)}) + H(X_i|X^{i-1}), \end{aligned}$$

where (i) uses the chain rule for entropy, and (ii) follows from the fact that conditioning reduces entropy. The above inequality is true for all values of $i \in [n]$, and hence summing them up, we get

$$nH(X^n) \leq \sum_{i=1}^{n} H(X^{(i)}) + \sum_{i=1}^{n} H(X_i|X^{i-1}) \overset{(iii)}{=} \sum_{i=1}^{n} H(X^{(i)}) + H(X^n),$$

where (iii) again follows from the chain rule for entropy. Subtracting $H(X^n)$ from both sides, and dividing by $n-1$ leads to the required result. □

Han's inequality has several applications in combinatorics and concentration of measure. We illustrate this with a simple result about the *density* of the subgraphs of binary hypercube.

Proposition 23. *Let $V = \{-1, 1\}^n$ denote the vertices of a binary hypercube in n dimensions, and let $E(V) = \{(x, x') : x, x' \in V,\ d_H(x, x') = 1\}$ denote the set of its edges (represented via unordered pairs of vertices). Note that $|V| = 2^n$, and $|E(V)| = n2^{n-1} = 2^{n-1}\log|V|$. Let A denote any subset of V, and $E(A)$ denote the edges between vertices in A. Then, we have*

$$|E(A)| \leq \frac{|A|}{2} \log|A|.$$

Proof. For the given subset $A \subset V = \{-1, 1\}^n$, let $X^n = (X_1, \ldots, X_n)$ be a random variable distributed uniformly over A. Then, by chain rule for entropy, we have the following for any $i \in [n]$:

$$H(X^n) - H(X^{(i)}) = H(X_i|X^{(i)}) = -\sum_{x^n \in A} p(x^n) \log p(x_i|x^{(i)}).$$

Given an $x^n \in A$, let $\bar{x}^{(i)}$ denote $(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n)$: the vector with the i^{th} element flipped. Then, we have

$$p(x_i|x^{(i)}) = 1 - \frac{1}{2}\mathbf{1}_{\bar{x}^{(i)} \in A}.$$

In other words, $\log(p(x_i|x^{(i)}) = \log(2) = 1$ if $\bar{x}^{(i)} \in A$, and 0 otherwise. Plugging this into the above equation, we get

$$H(X^n) - H(X^{(i)}) = \frac{1}{|A|} \sum_{x^n \in A} \mathbf{1}_{\bar{x}^{(i)} \in A},$$

which on summing over $i \in [n]$, gives us

$$\begin{aligned} \sum_{i=1}^{n} H(X^n) - H(X^{(i)}) &= \frac{1}{|A|} \sum_{x^n \in A} \sum_{i=1}^{n} \mathbf{1}_{\bar{x}^{(i)} \in A} \\ &= \frac{1}{|A|} \sum_{x^n \in A} |\{x' \in A : d_H(x, x') = 1\}| = \frac{2|E|}{|A|}. \end{aligned}$$

Finally, from an application of Han's inequality, we get

$$\frac{2|E(A)|}{|A|} = \sum_{i=1}^{n} H(X^n) - H(X^{(i)} = nH(X^n) - \sum_{i=1}^{n} H(X^{(i)}) \leq H(X^n) = \log(|A|).$$

This completes the proof. □

1.3.2 Convexity/Concavity

We begin with a simple consequence of the convexity of the mapping $x \mapsto x \log x$, that will be useful in establishing some properties of information measures.

Proposition 24. *Let $\{a_i, b_i : 1 \leq i \leq m\}$ denote non-negative real numbers (not all of which are zero), and define $A_m = \sum_{i=1}^m a_i$, and $B_m = \sum_{i=1}^m b_i$. Then, we have*

$$A_m \log\left(\frac{A_m}{B_m}\right) \leq \sum_{i=1}^{m} a_i \log\left(\frac{a_i}{b_i}\right).$$

The equality occurs if and only if a_i/b_i is a constant for all $i \in [m]$.

Proof. Introduce the terms $\tilde{a}_i = a_i/A_m$, and $\tilde{b}_i = b_i/B_m$, and note that $(\tilde{a}_1, \ldots, \tilde{a}_m)$, and $(\tilde{b}_1, \ldots, \tilde{b}_m)$ represent two probability distributions on $[m]$. Denote these probability distributions by P_a and P_b respectively. Then, we have

$$\begin{aligned}
\sum_{i=1}^{n} a_i \log(a_i/b_i) &= A_m \left(\sum_{i=1}^{m} \tilde{a}_i \log(\tilde{a}_i/\tilde{b}_i) + \log\left(\frac{A_m}{B_m}\right)\right) \\
&= A_m \log\left(\frac{A_m}{B_m}\right) + A_m \left(\sum_{i=1}^{m} \tilde{a}_i \log\left(\tilde{a}_i/\tilde{b}_i\right)\right) \\
&= A_m \log\left(\frac{A_m}{B_m}\right) + A_m D_{\mathrm{kl}}(P_a \parallel P_b) \\
&\geq A_m \log\left(\frac{A_m}{B_m}\right).
\end{aligned}$$

The inequality follows from the non-negativity of relative entropy, and the fact that $A_m \geq 0$ by assumption.

The equality occurs iff the relative entropy between P_a and P_b is zero:

$$D_{\mathrm{kl}}(P_a \parallel P_b) = 0 \;\Leftrightarrow\; P_a = P_b \;\Leftrightarrow\; \frac{a_i}{b_i} = \text{constant, for all } i \in [m].$$

□

As an application of the log-sum inequality, we can establish the convexity of relative entropy.

Theorem 25. *Let P_1, P_2, Q_1, Q_2 be distributions over $\mathcal{X}$. For any $\lambda \in [0,1]$, define $P_\lambda = \lambda P_1 + \bar{\lambda} P_2$, and $Q_\lambda = \lambda Q_1 + \bar{\lambda} Q_2$. Then, we have*

$$D_{kl}(P_\lambda \parallel Q_\lambda) \leq \lambda D_{kl}(P_1 \parallel Q_1) + \bar{\lambda} D_{kl}(P_2 \parallel Q_2).$$

Proof. For $\lambda \in \{0,1\}$, the result holds trivially. So we consider the case of $\lambda \in (0,1)$.

The relative entropy between P_λ and Q_λ is equal to

$$D_{\mathrm{kl}}(P_\lambda \parallel Q_\lambda) = \sum_{x \in \mathcal{X}} \lambda p_1(x) + \bar{\lambda} p_2(x) \log\left(\frac{\lambda p_1(x) + \bar{\lambda} p_2(x)}{\lambda q_1(x) + \bar{\lambda} q_2(x)}\right).$$

We now apply the log-sum inequality (Proposition 24) to each term in the summation. In particular, for a fixed $x \in \mathcal{X}$, define

$$a_1 = \lambda p_1(x), \quad a_2 = \bar{\lambda} p_2(x), \quad b_1 = \lambda q_1(x), \quad \text{and} \quad b_2 = \bar{\lambda} q_2(x).$$

An application of Proposition 24 implies that

$$(a_1 + a_2) \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \leq a_1 \log(a_1/b_1) + a_2 \log(a_2/b_2) = \lambda p_1(x) \log\left(\frac{p_1(x)}{q_1(x)}\right) + \bar{\lambda} p_2(x) \log\left(\frac{p_2(x)}{q_2(x)}\right).$$

The equality above uses the fact that $\lambda \not\in \{0,1\}$. On summing this upper bound over all $x \in \mathcal{X}$, we get the required result. □

A simple consequence of Theorem 25 is that the entropy is a concave functional of the pmf.

Corollary 26. *Suppose P_1 and P_2 are two distributions on $\mathcal{X}$. Then, for any $\lambda \in [0,1]$, we have*

$$H(P_\lambda) \geq \lambda H(P_1) + \bar{\lambda} H(P_2).$$

Proof. We know that the entropy of P_1 (resp. P_2) can be defined in terms of the relative entropy between P_1 (resp. P_2) and the uniform distribution over $\mathcal{X}$ (denoted by U):

$$H(P_1) = \log(|\mathcal{X}|) - D_{\mathrm{kl}}(P_1 \parallel U), \quad \text{and} \quad H(P_2) = \log(|\mathcal{X}|) - D_{\mathrm{kl}}(P_2 \parallel U).$$

This implies that for any $\lambda \in [0,1]$, we have

$$\begin{aligned} \lambda H(P_1) + \bar{\lambda} H(P_2) &= \log(|\mathcal{X}|) - \left(\lambda D_{\mathrm{kl}}(P_1 \parallel U) + \bar{\lambda} D_{\mathrm{kl}}(P_2 \parallel U)\right) \\ &\leq \log(|\mathcal{X}|) - D_{\mathrm{kl}}(P_\lambda \parallel U) \\ &= H(P_\lambda), \end{aligned}$$

where the inequality follows from the convexity of relative entropy. □

Finally, we characterize the convexity/concavity of the mutual information.

Theorem 27. *For two random variables X and Y taking values on finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively,*

- *for a fixed conditional distributions $\{p_{Y|X}(\cdot|x) : x \in \mathcal{X}\}$, the mapping $p_X \to I(X;Y)$ is concave.*
- *for a fixed marginal p_X, the mapping $p_{Y|X} \to I(X;Y)$ is convex.*

Proof. For the first statement, note that if the conditional distribution (or the channel, or the Markov kernel) $p_{Y|X}$ is fixed, then p_Y is a linear function of p_X. To see this, if we think of p_X, p_Y as row vectors, and use K to represent the $|\mathcal{X}| \times |\mathcal{Y}|$ transition matrix corresponding to $p_{Y|X}$, then $p_Y = p_X K$. Now, we observe that

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

Since $H(Y)$ is a concave function of p_Y, which in turn is a linear function of p_X, we conclude that $H(Y)$ is a concave function of p_X. The second term is simply a linear function of p_X, and hence their difference, $I(X;Y)$, is a concave function of p_X, with $P_{Y|X}$ fixed.

For the second result, we state the mutual information in terms of relative entropy. In particular, we have

$$I(X;Y) = D_{\mathrm{kl}}(P_{XY} \parallel P_X \times P_Y) = D_{\mathrm{kl}}(P_X \parallel P_X) + D_{\mathrm{kl}}(P_{Y|X} \parallel P_Y|P_X).$$

We already know that relative entropy is convex in its arguments, and the result follows by noting that for two Markov kernels, K_1 and K_2, and a $\lambda \in [0,1]$, we have

$$p_Y^\lambda = \lambda(p_X K_1) + \bar{\lambda}(p_X K_2) = p_X \left(\lambda K_1 + \bar{\lambda} K_2\right).$$

In particular, we have

$$\begin{aligned} D_{\mathrm{kl}}(K^\lambda \parallel p_Y^\lambda | P_X) &= D_{\mathrm{kl}}\left(\lambda K_1 + \bar{\lambda} K_2 \parallel \lambda p_Y^1 + \bar{\lambda} p_Y^2 | P_X\right) \\ &\leq \lambda D_{\mathrm{kl}}(K_1 \parallel p_Y^1 | P_X) + \bar{\lambda} D_{\mathrm{kl}}(K_2 \parallel p_Y^2 | P_X), \end{aligned}$$

as required. □

Remark 28. We know that by definition $I(X;Y)$ is always upper bounded by $H(X)$, which is further upper bounded by $\log(|\mathcal{X}|)$. Now, for a fixed channel (i.e., transition kernel), the above result says that $I(X;Y)$ is a non-negative, concave function of the marginal p_X. Hence, the term

$$C = \max_{p_X} I(X;Y),$$

is well-defined and is called the *channel capacity* associated with the channel $\{p_{Y|X=x} : x \in \mathcal{X}\}$.

1.3.3 Data Processing Inequality.

In this section we look at a class of inequalities called that *data processing inequalities.* Informally, these results tell us that we cannot increase the information contained in a signal (Y) about another unknown signal (X) by processing Y.

Theorem 29. *Suppose $X \to Y \to Z$ form a Markov chain; that is, $X \perp Z|Y$. Then, we have*

$$I(X;Z) \leq I(X;Y), \quad \textit{and} \quad I(X;Z) \leq I(Y;Z).$$

Proof. We show the proof of the first inequality, since the second inequality can be proved with the same argument.

We begin by writing $I(X;Z)$ in two equivalent ways:

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

Since conditioning reduces entropy, we have $H(X|Z) \geq H(X|Z,Y)$. Furthermore, due to the Markov property, $X \perp Z|Y$, which means that $H(X|Z,Y) = H(X|Y)$. This implies that

$$I(X;Z) = H(X) - H(X|Z) \leq H(X) - H(X|Z,Y) = H(X) - H(X|Y) = I(X;Y).$$

Note that the above relation holds with an equality iff $H(X|Z) = H(X|Z,Y)$. Or in other words, $X \to Z \to Y$ form a Markov chain. □

An immediate corollary of the above result gives us a data processing inequality for entropy.

Corollary 30. *Suppose $X \to Y \to Z$. Then, we have*

$$H(X|Y) \leq H(Z|Y).$$

Proof. The result follows from Theorem 29 by expanding $I(X;Y$ and $I(X;Z)$ in terms of the entropies. That is,

$$I(X;Z) \leq I(X;Y) \;\Rightarrow\; H(X) - H(X|Z) \leq H(X) - H(X|Y) \;\Rightarrow\; H(X|Y) \leq H(X|Z),$$

as required. □

The above result simply says that the amount of residual uncertainty about X given Y is never larger than the residual uncertainty about X given Z; where Z is some (possibly random) transform of Y.

Finally we present a DPI for relative entropy.

Theorem 31. *Consider two distributions P_X and Q_X over the alphabet $\mathcal{X}$, and let $K_{Y|X}$ denote a transition probability matrix. Then, with $P_Y = P_X K_{Y|X}$, and $Q_Y = Q_X K_{Y|X}$, we have*

$$D_{kl}(P_X \parallel Q_X) \geq D_{kl}(P_Y \parallel Q_Y).$$

In particular, if Y is any deterministic function of X, then we have

$$D_{kl}(P_X \parallel Q_X) \geq D_{kl}(P_{f(X)} \parallel Q_{f(X)}).$$

Proof. This statement is a simple consequence of the chain rule, and nonnegativity of relative entropy. In particular, using the chain rule, we can expand the relative entropy in two ways. First, we get

$$\begin{aligned} D_{\mathrm{kl}}(P_{XY} \parallel Q_{XY}) &= D_{\mathrm{kl}}(P_X \parallel Q_X) + D_{\mathrm{kl}}(Q_{Y|X} \parallel Q_{Y|X}|P_X) \\ &= D_{\mathrm{kl}}(P_X \parallel Q_X) + D_{\mathrm{kl}}(K_{Y|X} \parallel K_{Y|X}|P_X) \\ &= D_{\mathrm{kl}}(P_X \parallel Q_X). \end{aligned} \tag{1.5}$$

Now, expanding it the other way, we get

$$\begin{aligned} D_{\mathrm{kl}}(P_{XY} \parallel Q_{XY}) &= D_{\mathrm{kl}}(P_Y \parallel Q_Y) + D_{\mathrm{kl}}(Q_{X|Y} \parallel Q_{X|Y}|P_Y) \\ &\geq D_{\mathrm{kl}}(P_Y \parallel Q_Y), \end{aligned} \tag{1.6}$$

where the inequality follows from the nonnegativity of $D_{\mathrm{kl}}(Q_{X|Y} \parallel Q_{X|Y}|P_Y)$. Combining (1.5) and (1.6), we get the required result. □

Application: Impossibility results in hypothesis testing. Consider a hypothesis testing problem with i.i.d. observations $X_1, X_2, \ldots, X_n \in \mathcal{X}$ drawn from a distribution P_X, with mean μ_X with the null and alternative hypotheses defined as follows:

$$H_0 : \mu_X = 0.5, \quad \text{versus} \quad H_1 : \mu_X \geq 0.5 + \Delta.$$

Suppose there exists a test $\Psi : \mathcal{X}^n \to [0,1]$, with both type-I and type-II errors controlled at a level $\alpha \in (0,1)$. That is, the following two statements are simultaneously true:

$$p_1 \coloneqq \mathbb{E}_{H_1}[\Psi(X^n)] \geq 1 - \alpha \coloneqq \bar{\alpha}, \quad \text{and} \quad p_0 \coloneqq \mathbb{E}_{H_0}[\Psi(X^n)] \leq \alpha, \tag{1.7}$$

for some $\alpha \in (0, 1/4]$. Then, with a straightforward application of the DPI for relative entropy, we can obtain a lower bound on the parameter Δ. In particular, suppose H_1 is true, and the true distribution P_X has mean $\mu \geq 0.5 + \Delta$. Let P_0 denote the null distribution; that is Bernoulli with mean 0.5

$$\begin{aligned} nD_{\mathrm{kl}}\left(P_X \parallel P_0\right) = D_{\mathrm{kl}}(P_X^n \parallel P_0^n) &\overset{(a)}{\geq} D_{\mathrm{kl}}\left(\mathbb{E}_{P_X^n}[\Psi(X^n)] \parallel \mathbb{E}_{P_0^n}[\Psi(X^n)]\right) \\ &\geq D_{\mathrm{kl}}\left(\bar{\alpha} \parallel \alpha\right) = \bar{\alpha}\log(\bar{\alpha}/\alpha) + \alpha\log(\alpha/\bar{\alpha}) \\ &= (1-2\alpha)\log(\bar{\alpha}/\alpha) \geq \frac{1}{2}\log\left(\frac{1}{2\alpha}\right). \end{aligned}$$

The inequality (a) follows from an application of the DPI for relative entropy. The above chain of inequality says that if there exists a test Ψ satisfying (1.7), then the null and alternatives must be separated in relative entropy by at least $\log(1/2\alpha)/2n$. In other words,

$$\begin{aligned} &D^*(\Delta) \coloneqq \inf\{D_{\mathrm{kl}}(P_X \parallel P_0) : \mu_X \geq 1/2 + \Delta\} \geq A_n \coloneqq \frac{1}{2n}\log\left(\frac{1}{2\alpha}\right) \\ \Rightarrow\ &\Delta \geq \left(D^*\right)^{-1}(A_n) \coloneqq \min\{\Delta' > 0 : D^*(\Delta') \geq A_n\}. \end{aligned}$$

1.3.4 Fano's inequality.

Consider the Markov chain $X \to Y \to \hat{X}$. Here, X might denote some signal that is transmitted through a noisy channel, Y is the output of the noisy channel, and $\hat{X}$ might denote an estimate of X constructed on the basis of Y by the decoder. Assume that both X and $\hat{X}$ lie in some finite set $\mathcal{X}$. If $p_e = \mathbb{P}(\hat{X} \neq X)$ denotes the probability of error, then it is intuitive to expect p_e to depend on the amount of residual uncertainty about X given that we know Y. Fano's inequality is one way of formalizing this intuition.

Proposition 32. *For any decoder $\hat{X}$, we have the following:*

$$h_2(p_e) + p_e\log\left(|\mathcal{X}| - 1\right) \geq H(X|\hat{X}) \geq H(X|Y).$$

Remark 33. Note that it suffices to prove the first inequality, as the second follows from an application of data-processing inequality for entropy.

Proof. Introduce the Bernoulli random variable $Z = \mathbf{1}_{X \neq \hat{X}}$. Note that $\mathbb{P}(Z=1) = \mathbb{P}(X \neq \hat{X}) = p_e$. Then, consider the following:

$$H(X, Z|\hat{X}) = H(X|\hat{X}) + H(Z|X, \hat{X}) = H(X|\hat{X}), \tag{1.8}$$

where we used the fact that Z is a deterministic function of X and $\hat{X}$, which implies that $H(Z|X,\hat{X}) = 0$. Now, expanding $H(X, Z|\hat{X})$ another way, we get

$$\begin{aligned} H(X, Z|\hat{X}) &= H(Z|\hat{X}) + H(X|\hat{X}, Z) \overset{(i)}{\leq} H(Z) + H(X|\hat{X}, Z) \\ &= H(p_e) + \mathbb{P}(Z=1) H(X|\hat{X}, Z=1) + \mathbb{P}(Z=0) H(X|\hat{X}, Z=0) \\ &\overset{(ii)}{=} H(p_e) + \mathbb{P}(Z=1) H(X|\hat{X}, Z=1) + 0 \\ &\overset{(iii)}{\leq} H(p_e) + p_e \log\left(|\mathcal{X}| - 1\right). \end{aligned} \tag{1.9}$$

In the above display,
(i) uses the fact that conditioning reduces entropy,
(ii) uses the fact that when $Z = 0$, then X is equal to $\hat{X}$,
(iii) uses the fact that when $Z = 1$, X can take one of $|\mathcal{X}| - 1$ values in $\mathcal{X}$ not equal to $\hat{X}$.
The result then follows by combining (1.8) with (1.9). □

A simple corollary of the above inequality is often useful in statistical applications.

Corollary 34. *Suppose X is uniformly distributed over $\mathcal{X}$. Then, we have*

$$p_e \geq 1 - \frac{I(X;\hat{X}) - 1}{\log(|\mathcal{X}|)}.$$

Proof. The starting point of this result is the standard version of Fano's inequality:

$$-H(X|\hat{X}) \geq -h_2(p_e) - p_e \log(|\mathcal{X}|) \quad \Rightarrow \quad H(X) - H(X|\hat{X}) \geq -h_2(p_e) + (1-p_e)H(X).$$

Noting that $h_2(p_e) \leq 1$ (bits), on rearranging the above, we get

$$p_e H(X) \geq H(X) - I(X;\hat{X}) - 1.$$

We get the required statement by dividing both sides by $H(X) = \log(|\mathcal{X}|)$. □

The two inequalities above give us a bound on the probability of error under an exact recovery criterion: that is, $p_e = \mathbb{P}(\hat{X} \neq X)$. We now present a simple generalization for the case of approximate recovery, when the domain $\mathcal{X}$ is endowed with a distance measure d.

Corollary 35. *For a real number $t > 0$, introduce $p_t = \mathbb{P}\big(d(\hat{X}, X) > t\big)$, and define*

$$N_{\max}(t) = \max_{x \in \mathcal{X}} |B(x,t)|,$$

where $B(x,t) = \{x' \in \mathcal{X} : d(x,x') \leq t\}$ denotes the closed ball of radius t around x. Suppose $X \sim$ Uniform($\mathcal{X}$), and $X \to Y \to \hat{X}$ form a Markov chain. Then, we have

$$p_t \geq 1 - \frac{I(X,\hat{X}) - 1}{\log\left(\frac{|X|}{N_{\max}(t)}\right)}.$$

Proof. This result follows form the same general idea. We define $Z = \mathbf{1}_{d(X,\hat{X})>t}$, and note that $Z \sim \text{Bernoulli}(p_t)$.

$$H(X, Z|\hat{X}) = H(X|\hat{X}) + H(Z|X, \hat{X}) = H(Z|\hat{X}) + H(X|Z, \hat{X}).$$

As before, $H(Z|X, \hat{X}) = 0$. Furthermore, we have

$$\begin{aligned} H(Z|\hat{X}) + H(X|Z, \hat{X}) &\leq h_2(p_t)p_t H(X|Z = 1, \hat{X}) + (1 - p_t)H(X|Z = 0, \hat{X}) \\ &\leq h_2(p_t) + p_t \log(|\mathcal{X}|) + (1 - p_t) \log N_{\max}(t) \\ &= h_2(p_t) + p_t \log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) + \log(N_{\max}(t)). \end{aligned}$$

Plugging this back we get

$$H(X|\hat{X}) \leq h_2(p_t) + p_t \log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) + \log(N_{\max}(t)),$$

which implies

$$\begin{aligned} p_t &\geq \Big(H(X|\hat{X}) - h_2(p_t) - \log(N_{\max}(t))\Big) \,/\log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) \\ &= \Big(H(X|\hat{X}) - H(X) + H(X) - h_2(p_t) - \log(N_{\max}(t))\Big) \,/\log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) \\ &= \Big(H(X) - \log(N_{\max}(t)) - I(X;\hat{X}) - h_2(p_t)\Big) \,/\log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) \\ &= \Big(\log(|\mathcal{X}|) - \log(N_{\max}(t)) - I(X;\hat{X}) - h_2(p_t)\Big) \,/\log\left(\frac{\log|\mathcal{X}|}{N_{\max}(t)}\right) \\ &= 1 - \frac{I(X;\hat{X}) + h_2(p_e)}{\log\left(|\mathcal{X}|/N_{\max}(t)\right)} \geq 1 - \frac{I(X;\hat{X}) + 1}{\log\left(|\mathcal{X}|/N_{\max}(t)\right)} \end{aligned}$$

□

Chapter 2

Information measures for general distributions

We now extend the definition of the information measures beyond discrete distributions to more general spaces. First, we consider the case of continuous distributions in $\mathbb{R}^d$, and introduce the analogs of entropy, relative entropy, and mutual information. We study their properties and compare them with their discrete counterparts. Next, we define relative entropy and mutual information for general probability spaces, and establish two variational definitions. Finally, we introduce a class of information measures that generalize relative entropy, called the f-divergences, and study some of their properties.

2.1 Continuous Distributions

First we focus on the case of continuous distributions on $\mathcal{X} = \mathbb{R}^d$. That is, we focus on distributions P_X which admit a density, denoted by f_X, with respect to the Lebesgue measure on $\mathcal{X}$. This implies that for any measurable $E \subset \mathcal{X}$, we have $P_X = \int_{\mathcal{X}} \mathbf{1}_E(x) f_X(x)dx = \int_E f_X(x)dx$. For distributions (X, Y) with joint density f_{XY}, we also assume the existence of conditional densities $f_{Y|X}$ and $f_{X|Y}$ satisfying

$$P_{XY}(E) = \int_{\mathcal{X}} f_X(x)dx \int_{\mathcal{Y}} f_{Y|X}(y)\mathbf{1}_E(x,y)dy = \int_{\mathcal{Y}} f_Y(y)dy \int_{\mathcal{X}} f_{X|Y}(x)\mathbf{1}_E(x,y)dx.$$

For such distributions, we can define the continuous analogs of entropy, relative entropy, and mutual information, as well as their conditional variants.

Definition 36 (Differential Entropy)**.** The differential entropy of a continuous random variable X, with density f_X, is defined as

$$h(X) \equiv h(f_X) = -\int_S f_X(x)\log(f_X(x))dx,$$

where $S = \{x \in \mathcal{X} : f(x) > 0\}$ is the support of f_X. Similarly, conditional entropy of Y given X is defined as follows (where the integrals are over the appropriate supports):

$$h(Y|X) = -\int f_X(x)dx \int f_{Y|X}(y|x)\log\left(f_{Y|X}(y|x)\right)dy.$$

As in the discrete case, we can write the joint differential entropy $h(X, Y)$ as

$$h(X,Y) = h(X) - h(Y|X) = h(Y) - h(Y|X). \tag{2.1}$$

The definition of joint entropy in (2.1) implicitly assumes that at least of the terms is not infinite. We now evaluate the differential entropy for some common distributions.

Example 37 (Uniform distribution)**.** Suppose $\mathcal{X} = [a, b]$, and X is a the uniformly distributed random variable over $\mathcal{X}$ with density $f_X(x) = 1/(b-a)$ for all $x \in \mathcal{X}$. Then, the differential entropy of X is equal to

$$h(X) = \int_{\mathcal{X}} \frac{1}{b-a} \log(b-a) dx = \log(b-a).$$

Thus, $h(X) = 0$ for $(b-a) = 1$, $h(X) < 0$ for $(b-a) < 1$, and $h(X) > 0$ for $(b-a) > 1$. This is in contrast with entropy defined for discrete distributions, which is always non-negative, and equal to 0 only for Dirac distributions.

Example 38 (Multivariate Gaussian)**.** Consider a multivariate Gaussian random variable (X) over $\mathcal{X} = \mathbb{R}^d$, with mean μ and covariance matrix K. The density of this random variable is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |K|}} \exp\left(-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)\right).$$

Then, the differential entropy of X is equal to

$$\begin{aligned} h(X) &= \log\left(\sqrt{(2\pi)^d |K|}\right) + \frac{1}{2}\int_{\mathcal{X}} (x-\mu)^T K^{-1}(x-\mu) f_X(x) dx \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\mathbb{E}\left[(X-\mu)^T K^{-1}(X-\mu)\right] \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \mathbb{E}\left[(X_i-\mu_i)K_{ij}^{-1}(X_j-\mu_j)\right] \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} K_{ij}^{-1}\mathbb{E}\left[(X_i-\mu_i)(X_j-\mu_j)\right] \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} K_{ij}^{-1}K_{ij} \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} K_{ij}^{-1}K_{ji} \\ &= \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{1}{2}\sum_{i=1}^{d}(I_d)_{ii} = \frac{1}{2}\log\left((2\pi)^d |K|\right) + \frac{d}{2} \\ &= \frac{1}{2}\log\left((2\pi e)^d |K|\right) \text{ nats/bits.} \end{aligned}$$

A simple consequence of this is that for univariate Gaussian with mean μ, and variance σ^2, the differential entropy is equal to $(1/2)\log(2\pi e\sigma^2)$ nats/bits.

Remark 39. Note that in both the examples considered above, the differential entropy is "translation invariant": that is, the differential entropy of X and $X + c$ for any $c \in \mathcal{X}$ is the same. This property is true generally for differential entropy, beyond the two examples above, as we will see later.

We now introduce the definition of relative entropy for continuous distributions.

Definition 40. Suppose P_X and P_Y are two continuous distributions on $\mathcal{X}$, with densities f_X and f_Y respectively. Then, the relative entropy between them is defined as (with S_X denoting the support of P_X):

$$D_{\text{kl}}(P_X \parallel P_Y) = \begin{cases} \int_{S_X} f_X(x) \log\left(\frac{f_X(x)}{f_Y(x)}\right) dx, & \text{if } \{f_Y = 0\} \subset \{f_X = 0\}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Similarly, let P_{XY} (with density f_{XY}) and Q_{XY} (with density g_{XY}) denote two continuous joint distributions over $\mathcal{X} \times \mathcal{Y}$. Then the conditional relative entropy between $P_{Y|X}$ and $Q_{Y|X}$ is defined as

$$D_{\text{kl}}(P_{Y|X} \parallel Q_{Y|X}|P_X) = \int f_X(x)dx \int f_{Y|X}(y|x) \log\left(\frac{f_{Y|X}(y|x)}{g_{Y|X}(y|x)}\right) dy.$$

Example 41. The relative entropy between two univariate Gaussians, $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$ is equal to

$$\begin{aligned}
D_{\mathrm{kl}}(P \parallel Q) &= \mathbb{E}_P\left[\log(p(X)/q(X))\right] = \mathbb{E}_P\left[\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{(X-\mu_2)^2}{\sigma_2^2} - \frac{(X-\mu_1)^2}{\sigma_1^2}\right)\right] \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\mathbb{E}_P\left[\left(\frac{(X-\mu_2)^2}{\sigma_2^2} - \frac{(X-\mu_1)^2}{\sigma_1^2}\right)\right] \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\sigma_2^2 + (\mu_1-\mu_2)^2}{\sigma_2^2} - 1\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{(\mu_1-\mu_2)^2}{2}.
\end{aligned}$$

Thus, if $\sigma_1 = \sigma_2$, the relative entropy is proportional to the difference in mean squared.

Example 42 (Gaussian-Cauchy)**.** Note that the relative entropy can be infinite even if the distributions are absolutely continuous. For example, consider the relative entropy between P and Q, when P has a Cauchy distribution with density $f(x) = 1/\left(\pi(x^2+1)\right)$, and Q has a Gaussian distribution with density $g(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$. Clearly, $P \ll Q$ and $Q \ll P$. However,

$$D_{\mathrm{kl}}(P \parallel Q) = \mathbb{E}_P\left[\log(p(X)/q(X)\right] \asymp -h(P) + \mathbb{E}_P[X^2] = \infty,$$

since $h(P)$ can be shown to be finite, while the second moment of Cauchy distribution is infinite.

Finally, we can now introduce the definition of mutual information for continuous random variables.

Definition 43. The mutual information between two continuous random variables X and Y is defined as

$$I(X;Y) = \int_{S_{XY}} f_{XY}(x,y)\log\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)dx\,dy,$$

where $S_{XY} = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : f_X(x,y) > 0\}$ is the support of the joint density of (X,Y). It is easy to check that $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$.

2.1.1 Connections to discrete measures via quantization

Example 44. Let U denote the random variable with uniform distribution on $[0,a]$. For any $\Delta > 0$, let U_Δ denote the discrete quantization of U: that is, U_Δ takes values in the set $\{i\Delta : 1 \le i \le a/\Delta\}$. Then, we have the following relation:

$$H(U_\Delta) = \log(a/\Delta) = \log(a) - \log(\Delta) = h(U) - \log(\Delta).$$

Or in other words, for any $\Delta > 0$ (such that a/Δ is an integer), we have

$$H(U_\Delta) + \log(\Delta) = h(U).$$

We now show that the same relation between $H(\cdot)$ and $h(\cdot)$ holds more generally for continuous distributions.

Definition 45. Let X denote a real-valued random variable with continuous density f_X. For any $\Delta > 0$, partition $\mathcal{X} = \mathbb{R}$ into a (countable) grid of size Δ, consisting of sets

$$E_i = [i\Delta, (i+1)\Delta), \quad \text{for } -\infty \le i \le \infty.$$

Then, for any i, due to the mean-value theorem, there exists an $x_i \in E_i$, such that $f_X(x_i)\Delta = \int_{E_i} f_X(x)dx \coloneqq p_i$. We define the Δ-quantized version of X, denoted by X_Δ, as

$$X_\Delta = x_i, \quad \text{with probability } p_i, \quad \text{for } i \in \mathbb{Z}.$$

Theorem 46. *Consider the case of $\mathcal{X} = \mathbb{R}$, and let X be a continuous distribution with (a Riemann integrable) density function f_X. Then, we have the following:*

$$\lim_{\Delta \downarrow 0} H(X_\Delta) + \log(\Delta) = h(X) = h(f_X),$$

where X_Δ is the Δ-quantized version of X (Definition 45).

Proof. The proof follows directly from the definition of discrete entropy. In particular,

$$\begin{aligned}
-H(X_\Delta) &= \sum_{i\in\mathbb{Z}} p_i \log p_i = \sum_{i\in\mathbb{Z}} f_X(x_i)\Delta\left(\log(f_X(x_i)) + \log(\Delta)\right) \\
&= \sum_{i\in\mathbb{Z}} f_X(x_i)\Delta \log(f_X(x_i)) + \log(\Delta)\sum_{i\in\mathbb{Z}} p_i \\
&= \sum_{i\in\mathbb{Z}} f_X(x_i)\Delta \log(f_X(x_i)) + \log(\Delta),
\end{aligned}$$

where the last equality simply uses the fact that $\sum_{i\in\mathbb{Z}} p_i = 1$. On rearranging the above relation, we get

$$H(X_\Delta) + \log\Delta = -\sum_{i\in\mathbb{Z}} f_X(x_i)\Delta\log(f_X(x_i)).$$

Since the term on the right is the Riemann sum for the integral $-\int_{\mathbb{R}} f_X(x)\log(f_X(x))dx$ that defines the differential entropy of X, we get the required result by taking the limit $\Delta \downarrow 0$:

$$\lim_{\Delta\downarrow 0} H(X_\Delta) + \log\Delta = h(f_X).$$

□

Remark 47. One interpretation of this above result is in terms of the number of bits needed to learn a Δ-approximate version of a continuous random variable X. That is, to approximate a continuous random variable X, with an approximation error Δ, we roughly need $\log(1/\Delta) + h(f_X)$ bits. For example, going back to our uniform example, to represent a uniform $[0, a]$ random variable with up to Δ error, we need $\log(a/\Delta)$ bits (exactly).

Remark 48. Another interpretation of the above result is via the AEP for the discrete and continuous entropies. For concreteness, let X be a uniformly distributed over $\mathcal{X} = [0, a]$ random variable, and X_Δ is its Δ-quantization. Then, we have the following two statements:

- AEP for the differential entropy says that (with probability almost 1), the n i.i.d. realizations of X are concentrated in a volume of $\approx 2^{nh(X)}$ within the space $\mathcal{X}^n$.
- AEP for discrete entropy says that (with probability almost 1) there are $\approx 2^{hH(X_\Delta)}$ equiprobable sequences of length n in the quantized space $\mathcal{X}_\Delta^n$.

Each point in $\mathcal{X}_\Delta^n$ denotes a unique cube in $\mathcal{X}^n$, of volume Δ^n. Hence, the total volume covered by the $2^{nH(X_\Delta)}$ sequences is $\approx \Delta^n 2^{nH(X_\Delta)}$. The result above says that the two volumes are approximately equal: that is,

$$\Delta^n 2^{nH(X_\Delta)} \approx 2^{nh(X)}, \qquad \text{or} \qquad \log(\Delta) + H(X_\Delta) \approx h(X).$$

Unlike the entropy, the behavior of relative entropy and mutual information is stable under quantization. In fact, through informal arguments, we can see that the continuous relative entropy (and thus mutual information as well) can be seen as the limits of their quantized versions, as $\Delta \to 0$.

2.1.2 Properties

We begin by noting the nonnegativity of relative entropy for continuous distributions.

Proposition 49. *For two $\mathcal{X} = \mathbb{R}^d$-valued distributions P and Q with densities f and g respectively, we have*

$$D_{kl}(P \parallel Q) \equiv D_{kl}(f \parallel g) \geq 0,$$

with equality if and only if $f = g$ almost everywhere. As consequence, we also have for any two continuous random variables X and Y:

- *$I(X;Y) \geq 0$, with equality if and only if $X \perp Y$.*

- *$h(X|Y) \leq h(X)$, with equality if and only if $X \perp Y$.*

Proof. The proof of the nonnegativity of D_{kl} follows directly from the convexity of the mapping $\varphi(x) = x \log x$. First, note that if $P \not\ll Q$, then $D_{\mathrm{kl}}(P \parallel Q) = \infty$, and there is nothing to prove. Hence, we can assume that $P \ll Q$, and $S_f = \{x : f(x) > 0\} \subset S_g = \{x : g(x) > 0\}$. Then, we have

$$\begin{aligned} D_{\mathrm{kl}}(f \parallel g) &= \int_{S_f} g(x)\varphi\left(\frac{f(x)}{g(x)}\right) dx = \inf_{S_g} g(x)\varphi\left(\frac{f(x)}{g(x)}\right) dx \\ &\geq \varphi\left(\int_{S_g} g(x)\frac{f(x)}{g(x)} dx\right) = \varphi(1) = 0. \end{aligned}$$

Since the $\geq$ above follows from Jensen's inequality, it holds with equality if and only if $f(x)/g(x)$ is a constant almost surely. Since both of them are densities, the constant must be 1 and thus the equality holds when $f = g$ almost surely. □

We now look at some properties of differential entropy.

Proposition 50. *The differential entropy of continuous random variables satisfies the following properties:*

(a) Chain rule: $h(X_1, \ldots, X_n) = \sum_{i=1}^n h(X_i|X^{i-1})$.

(b) Translation invariance: $h(X + c) = h(X)$ for all $c \in \mathcal{X}$.

(c) Effect of scaling: if $\mathcal{X} = \mathbb{R}$, then $h(aX) = h(X) + \log(|a|)$. More generally, for vector valued X, we have $h(AX) = h(X) + \log(|det(A)|)$, where A is an invertible matrix.

Proof. (a) Follows directly from the definition.

(b) Let $Y = X + c$. Then, we have $f_Y(x) = f_X(x - c)$, and hence we have

$$h(Y) = \int_{c+S_X} f_X(x-c)\log(1/f_X(x-c))dx = \int_{S_X} f_X(x)\log(1/f_X(fX))dx = h(X).$$

(c) We prove the general multivariate case. Let $Y = AX$, and by the change of variable formula, we have

$$g(y) = \frac{1}{|A|} f(A^{-1}y).$$

Then, we have the following

$$\begin{aligned} h(AX) = h(Y) = -\int g(y)\log(g(y)) &= -\frac{1}{|A|}\int f(A^{-1}y)\left(\log(f(A^{-1}y) - \log(|A|)\right) \\ &= \log(|A|) - \frac{1}{|A|}\int f(A^{-1}y)\log\left(f(A^{-1}y)\right) dy \\ &\overset{(i)}{=} \log(|A|) - \int f(x)\log\left(f(x)\right) dx = h(X) + \log(|A|). \end{aligned}$$

In the equality (i) above, we used the change of variable $x = A^{-1}y$, which implies $dx = |\det(A^{-1})|dy = (1/|A|)dy$.

□

Remark 51. Note that despite the scaling operation being a bijection, the differential entropy of the resulting variable (i.e., $Y = AX$) is different from the original (X). This is unlike the discrete entropy which did not change under such operations.

We end this section with a result about a characterization of Gaussian in terms of the maximum entropy achieving distribution among the class of all distributions with finite second moment.

Proposition 52. *Let $\mathcal{P}_2(\mathbb{R})$ denote the class of continuous probability distributions on $\mathbb{R}$ with zero mean and with second moment upper bounded by $b < \infty$. Then, the distribution from $\mathcal{P}_2$ with the largest differential entropy is $N(0, b)$.*

Proof. Without loss of generality, we assume that $b = 1$.

Suppose P be any distribution in $\mathcal{P}_2$, with density f, and Let ϕ denote the density of the standard normal random variable. Then, by the nonnegativity of relative entropy, we have

$$\begin{aligned}
0 \leq D_{\mathrm{kl}}(f \parallel \phi) &= \int_{\mathbb{R}} f(x)\big(\log(f(x)) - \log(\phi(x))\big)dx \\
&= -\int_{\mathbb{R}} f(x)\log(\phi(x))dx + \int_{\mathbb{R}} f(x)\log(f(x))dx \\
&= -\int_{\mathbb{R}} \phi(x)\log(\phi(x))dx - h(f) \\
&= h(\phi) - h(f).
\end{aligned}$$

This completes the proof. □

Example 53 (Capacity of Gaussian Channel)**.** Suppose $\mathcal{X} = \mathbb{R}$, and consider a power-limited Gaussian channel, which takes in a $\mathcal{X}$-valued input X, and perturbs it with an additive Gaussian noise $Z \sim N(0, \sigma_N^2)$ that is independent of X. The output of the channel is denoted by $Y = X + Z$. Suppose we have an energy constraint $\mathbb{E}[X^2] \leq \sigma_P^2$. Then, the capacity of this channel is equal to

$$C = \sup_{f_X : \mathbb{E}[X^2] \leq \sigma_P^2} I(X;Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_P^2}{\sigma_N^2}\right) = \frac{1}{2}\log\left(1 + \mathrm{SNR}\right).$$

To see why this is true, not that for any arbitrary input X, we have

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

where the last equality uses the fact that $Z \perp X$. Hence, we have

$$I(X;Y) = h(Y) - \frac{1}{2}\log(2\pi e\sigma_N^2) \leq \frac{1}{2}\log(2\pi e\sigma_Y^2) - \frac{1}{2}\log(2\pi e\sigma_N^2),$$

where the inequality uses Proposition 52. Since $X \perp X$, the variance of Y (denoted by σ_Y^2) is equal to $\sigma_X^2 + \sigma_N^2$. Furthermore, under the constraint, the maximum value of σ_X^2 is equal to σ_P^2. Thus, we have

$$I(X;Y) \leq \frac{1}{2}\log\left(\frac{\sigma_N^2 + \sigma_P^2}{\sigma_N^2}\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_P^2}{\sigma_N^2}\right).$$

Since X was arbitrary, this also implies that $C \leq (1/2)\log(1 + \sigma_P^2/\sigma_N^2)$. Finally, note that the inequality holds with an equality for the input distribution $f_X = N(0, \sigma_P^2)$.

In the next section, we illustrate some interesting consequences of the simple properties of continuous information measures we obtained above.

2.1.3 Applications

We now look at some simple applications of the properties of continuous information measures.

Maximum Entropy distributions. Staying with the theme of maximum entropy distributions, we show a general method for characterizing the distributions achieving the optimal entropy under different types of constraints.

Let $\mathcal{F}$ denote a class of densities on $\mathcal{X} = \mathbb{R}$ satisfying m integral constraints:

- $f(x) \geq 0$, for all $x \in S$, with equality outside S.

- $\int_S f(x)dx = 1$.
- $\int_S f(x)r_i(x)dx = \alpha_i$, for $1 \leq i \leq m$.

Our next result characterizes the distribution from $\mathcal{F}$ achieving the maximum entropy.

Proposition 54. *The density $f^* \in \mathcal{F}$ achieving the maximum entropy is of the form $f^*(x) = \exp\left(\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)\right)$ for all $x \in S$, for some real-valued terms $\{\lambda_i : 0 \leq i \leq m\}$, chosen to ensure f^* satisfies the constraints.*

Proof. We will use a slightly informal "guess and verify" strategy to prove this result. In particular, note that entropy is a concave functional of the densities, and the domain is convex. Hence, we can write the Lagrangian as

$$L(f) = -\int f \log f + \lambda_0 \int f + \sum_{i=1}^m \int f r_i.$$

The proper way to proceed now is to take the functional derivative of J with respect to f. Instead, we present an informal argument. Suppose we write the integrals as approximate Riemann sums. That is, we consider a grid of width Δ partitioning the domain $\mathcal{X}$, and write

$$L(f) \approx -\sum_{j\in\mathbb{Z}} f(x_j)\log(f(x_j))\Delta + \lambda_0 \sum_{i\in\mathbb{Z}} f(x_j)\Delta + \sum_{i=1}^m \sum_{j\in\mathbb{Z}} f(x_j) r_i(x_j)\Delta.$$

Now, we can consider $L(f)$ as a function of the variables $\{y_j = f(x_j) : j \in \mathbb{Z}\}$. So taking the (usual) derivative of $L(f)$ with $y_j = f(x_j)$, and setting it to 0, we get

$$\frac{dL(f)}{dy_j} = \Delta\left(-\log(f(x_j)) - 1 + \lambda_0 + \sum_{i=1}^m f(x_j) r_i(x_i)\right) = 0.$$

The above equation suggests a solution of the form

$$f(x) = \exp\left(-1 + \lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)\right).$$

Now, constraints give us $(m+1)$ equations in $(m+1)$ unknowns $(\lambda_i)_{i=0}^m$, and we can solve this linear system to construct a candidate solution f^* of the above form, satisfying all the constraints.

Having informally justified the form of the candidate solution f^*, we now provide an information theoretic proof of its optimality. Let g denote any element of $\mathcal{F}$. Then, we have

$$h(g) = -\int_S g \log g = -\int_S g \log(g/f^*) - \int_S g \log f^* = -D_{\mathrm{kl}}(g \parallel f^*) - \int_S g \log f^*.$$

Since f^* is supported on the entire S, the above operation is valid. Due to the nonnegativity of relative entropy, the above result implies that

$$h(g) \leq -\int_S g \log f^*.$$

If the integral on the right side were with respect to the density f^*, we would have obtained the required inequality. Nevertheless, we exploit the specific form of f^* to show the required inequality:

$$\begin{aligned} -\int g \log f^* &= -\int g\left(\lambda_0 + \sum_{i=1}^m \lambda_i r_i\right) = -\lambda_0 \int_S g - \sum_{i=1}^m \lambda_i \int_S g r_i \\ &\overset{(i)}{=} -\lambda_0 \int_S f^* - \sum_{i=1}^m \lambda_i \int_S f^* r_i \\ &= -\int_S f^* \log f^* = h(f^*). \end{aligned}$$

The equality (i) follows from the fact that both g and f^* are elements of $\mathcal{F}$, and thus they satisfy the required constraints. □

Example 55. The previous result can be used to characterize the max-entropy distribution in various settings.

- $S = [a, b]$, and no other constraints. Then, the optimal distribution has the form $f^*(x) = e^{\lambda_0}$; that is, the max-ent distribution is the uniform distribution.

- Suppose $S = \mathbb{R}$, and the constraints are $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = \sigma^2$. Then, the optimal density has the form $f^*(x) = \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2\right)$. It is easy to verify that this is a Gaussian distribution with mean 0 and variance σ^2.

- Suppose $S = [0, \infty)$, and the only constraint is $\mathbb{E}[X] = \mu$. Then, the optimal distribution is of the form $f^*(x) = \exp(\lambda_0 + \lambda_1 x)$, which turns out to be the exponential distribution with parameter μ. That is, $f^*(x) = (1/\mu)e^{-x/\mu}$.

- Suppose $S = [0, \infty)$, and the constraints are $\mathbb{E}[X] = \alpha_1$ and $\mathbb{E}[\log X] = \alpha_2$. Then, the optimal distribution has the form $f^*(x) = \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 log(x)\right) = x^{\lambda_2} \exp\left(\lambda_0 + \lambda_1 x\right)$. This is the family of Gamma distributions.

- For $S = \mathbb{R}^d$, and the constraint that $E[X_i] = 0$ for all $i \in [d]$, and $\mathbb{E}[X_i X_j] = K_{ij}$, where $K \equiv [K_{ij}]$ is a positive semi-definite matrix, a similar argument shows that the distribution achieving the maximum entropy is $N(0, K)$.

Determinant Inequalities. The properties of continuous information measures, especially for Gaussian distributions, can be used to obtain elementary proofs of several nontrivial matrix inequalities.

Proposition 56. *The following statements are true.*

(a) Suppose K is an $n \times n$ positive semi-definite matrix. Then, we have

$$|K| \equiv |det(K)| \leq \prod_{i=1}^{n} K_{ii},$$

with equality if and only if K is a diagonal matrix.

(b) For any two psd matrices K_1 and K_2, and any $\lambda \in [0, 1]$, we have

$$|\lambda K_1 + \overline{\lambda} K_2| \geq |K_1|^{\lambda} |K_2|^{\overline{\lambda}}.$$

Or, in other words, the log-determinant function is concave on the space of positive semi definite matrices.

Proof. (a) Given the psd matrix K, consider the Gaussian random vector $(X_1, \ldots, X_n) \equiv X^n \sim N(0, K)$ on the domain $\mathcal{X} = \mathbb{R}^n$. Then, the chain rule for differential entropy implies that

$$h(X^n) = \sum_{i=1}^{n} h(X_i | X^{i-1}) \leq \sum_{i=1}^{n} h(X_i).$$

Since the marginals of X^n are also univariate Gaussians, with $X_i \sim N(0, K_{ii})$, we observe that

$$\frac{1}{2} \log\left((2\pi e)^n |K|\right) \leq \sum_{i=1}^{n} \frac{1}{2} \log\left(2\pi e |K_{ii}|\right) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log\left(\prod_{i=1}^{n} K_{ii}\right).$$

On cancelling $(n/2)\log(2\pi e)$ from both sides, and exponentiating, we get the required inequality.

Furthermore, we know that the subadditivity of differential entropy holds with equality if and only if all the $X_i's$ are independent; or in other words, the matrix K is a equal to $\text{diag}(K_{11}, \ldots, K_{nn})$.

(b) Let $X \sim N(0, K_1)$ and $Y \sim N(0, K_2)$ denote two independent multivariate Gaussian random variables on $\mathcal{X} = \mathbb{R}^n$. For a fixed $\lambda \in [0, 1]$, introduce a Bernoulli random variable $\Lambda \sim \text{Bernoulli}(\lambda)$ independent of X and Y, and use it to define the mixture $Z = \Lambda X + \overline{\Lambda} Y$.

We first observe that the covariance of Z is equal to $K_\lambda \coloneqq \lambda K_1 + \overline{\lambda} K_2$. To see this, note that since $\mathbb{E}[Z] = 0$, we have

$$\begin{aligned} \mathrm{Cov}(Z) = \mathbb{E}[ZZ^T] &= \mathbb{P}(\Lambda = 0)\mathbb{E}[YY^T|\Lambda = 0] + \mathbb{P}(\Lambda = 1)\mathbb{E}[XX^T|\Lambda = 1] \\ &= \overline{\lambda}\mathbb{E}[YY^T] + \lambda\mathbb{E}[XX^T] \\ &= \overline{\lambda}K_2 + \lambda K_1 = K_\lambda. \end{aligned} \tag{2.2}$$

Now, we use the fact that conditioning reduces entropy, to observe

$$h(Z|\Lambda) \leq h(Z) \leq h\left(N(0, K_\lambda)\right) = \frac{1}{2}\log\left((2\pi e)^n |K_\lambda|\right).$$

Finally, we calculate the conditional entropy of Z given Λ:

$$\begin{aligned} h(Z|\Lambda) &= \lambda h(Z|\Lambda = 1) + \overline{\lambda} h(Z|\lambda = 0) \\ &= \lambda h(X) + \overline{\lambda} h(Y) \\ &= \frac{\lambda}{2}\log\left((2\pi e)^n|K_1|\right) + \frac{\overline{\lambda}}{2}\log\left((2\pi e)^n|K_2|\right). \end{aligned} \tag{2.3}$$

Now, by combining (2.2) with (2.3), we get the required statement that

$$\lambda\log(|K_1|) + \overline{\lambda}\log(|K_2|) \leq \log(|K_\lambda|).$$

□

Estimation error. For the case of discrete distributions, Fano's inequality gave us a lower bound on the probability of error in a hypothesis testing problem in terms of the entropy of the observations. We now present two analogous results for the problem of estimation.

Proposition 57. *Consider the Markov chain $X \to Y \to \widehat{X}$, where $\widehat{X}$ is an estimate of the unknown random variable X. Then, we have*

$$\mathbb{E}[(X - \widehat{X}(Y))^2] \geq \frac{1}{2\pi e}2^{h(X|Y)}.$$

In the special case where the side information is not available, the above inequality reduces to

$$\mathbb{E}[(X - \widehat{X})^2] \geq \frac{1}{2\pi e}2^{h(X)}.$$

Proof. To prove this result, we begin with the observation that

$$\begin{aligned} \mathbb{E}[(X - \widehat{X}(Y))^2] &\geq \mathbb{E}[(X - E[X|Y])^2] = \mathbb{E}_Y\left[\mathbb{E}_{X|Y}\left[(X - E[X|Y])^2|Y\right]\right] \\ &= \mathbb{E}_Y[\mathbb{V}(X|Y)]. \end{aligned} \tag{2.4}$$

Now, observe that

$$\begin{aligned} h(X|Y) &= -\int_{\mathcal{Y}} f(y)dy\int_{\mathcal{X}} f(x|y)\log(f(x|y))dx \leq \int_{\mathcal{Y}} f(y)\frac{1}{2}\log\left(2\pi e\mathbb{V}(X|Y = y)\right) \\ &\overset{(i)}{\leq} \frac{1}{2}\log\left(2\pi e\,\mathbb{E}_Y[\mathbb{V}(X|Y)]\right), \end{aligned}$$

where (i) follows from an application of Jensen's inequality, along with the concavity of $x \mapsto \log(x)$. On rearranging the above inequality, we get

$$\frac{1}{2\pi e}2^{2h(X|Y)} \leq \mathbb{E}_Y[\mathbb{V}(X|Y)]. \tag{2.5}$$

Together, (2.4) and (2.5) imply the required result. □

Bounding the entropy of discrete distribution. We end with one final application of Proposition 52. For the case of discrete distributions with finite support, we know that the entropy is upper bounded by $\log(|\mathcal{X}|)$. However, this bound is not applicable when the support of a discrete random variable is countable. We now show how to use the max entropy continuous distribution to obtain a non-vacuous upper bound on the entropy of a discrete distribution with countable support.

Proposition 58. *Let $\mathcal{X} = \{x_i : i \in \mathbb{N}\}$ denote a countable alphabet, and let X be an $\mathcal{X}$-valued random variable with $\mathbb{P}(X = x_i) = p_i$ for $i \in \mathbb{N}$. Then, we have*

$$H(X) = H(p_1, p_2, \ldots) \leq \frac{1}{2} \log \left((2\pi e) \times \left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i \right)^2 + \frac{1}{12} \right) \right).$$

Proof. Introduce a new random variable Y, such that $\mathbb{P}(Y = i) = p_i$. Note that $H(Y) = H(X)$, since the discrete entropy is independent of the actual values taken by the random variable. Now, let U denote a Uniform $([0,1])$ random variable, and use it to define $Z = Y + U$. We can verify that Z is a continuous random variable with a density $f_Z(x) = p_i$, where $i = \lfloor x \rfloor$.

Now, the discrete entropy of Y is equal to

$$\begin{aligned} H(X) = H(Y) = -\sum_{i \in \mathbb{N}} p_i \log p_i &= \sum_{i \in \mathbb{N}} \int_i^{i+1} f_Z(x) \log\left(f_Z(x)\right) dx \\ &= -\int_{\mathbb{R}} f_Z(x) \log(f_Z(x)) dx \\ &= h(Z). \end{aligned}$$

Note that since X and U are independent, the variance of Z is equal to $\mathbb{V}(Y) + \mathbb{V}(U)$. Thus the differential entropy of $h(Z)$ is upper bounded by that of a zero mean Gaussian with the same variance. Hence, we have

$$\begin{aligned} H(X) = h(Z) &\leq \frac{1}{2} \log\left((2\pi e)\mathbb{V}(Z)\right), \quad \text{where} \\ \mathbb{V}(Z) = \mathbb{V}(Y) + \mathbb{V}(U) &= \sum_i i^2 p_i - \left(\sum_i i p_i \right)^2 + \frac{1}{12}. \end{aligned}$$

□

2.2 General Distributions*

In this section, we present the general definition of relative entropy. To state the definition, we need to recall the Radon-Nikodym theorem.

Fact 59. *Consider a measurable space $(\mathcal{X}, \mathcal{F})$ with two σ-finite measures P and Q. Suppose that P is absolutely continuous w.r.t. Q, denoted by $P \ll Q$, which means that $Q(E) = 0 \Rightarrow P(E) = 0$, for any $E \in \mathcal{F}$. Then, there exists a measurable function $f : \mathcal{X} \to [0, \infty)$, such that*

$$P(E) = \int_E f dQ, \quad \textit{for all } E \in \mathcal{F}.$$

The (not necessarily unique) function f is called the Radon-Nikodym derivative of P w.r.t. Q, and is also denoted as $\frac{dP}{dQ}$. If $P \ll Q$ and $Q \ll P$, we have

$$\frac{dP}{dQ} = \left(\frac{dQ}{dP} \right)^{-1}.$$

Finally, if $P \ll Q$ and $Q \ll R$, then we have

$$\frac{dP}{dR} = \frac{dP}{dQ} \times \frac{dQ}{dR}.$$

Since we deal with probability measures, that are finite and hence σ-finite, the only condition required for the existence of the Radon-Nikodym derivative is the absolute continuity. Using this, we have the following general definition of relative entropy.

Definition 60. For any two distributions P and Q, defined on a common measurable space $(\mathcal{X}, \mathcal{F})$, the relative entropy between P and Q is defined as

$$D_{\mathrm{kl}}(P \parallel Q) := \begin{cases} \mathbb{E}_Q \left[\frac{dP}{dQ}(X) \log \frac{dP}{dQ}(X) \right], & \text{if } P \ll Q, \\ +\infty, & \text{if } P \not\ll Q. \end{cases}$$

Fact 61. *An equivalent definition of the relative entropy between P and Q, is*

$$D_{kl}(P \parallel Q) := \begin{cases} \mathbb{E}_P \left[\log \frac{dP}{dQ}(X) \right], & \text{if } P \ll Q, \\ +\infty, & \text{if } P \not\ll Q. \end{cases}$$

See Lemma 2.4 of Polyanskiy and Wu for a proof.

Definition 62. Suppose $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ denote two measurable spaces. Then, a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ is a mapping $K : \mathcal{G} \times \mathcal{X} \to [0, 1]$, such that

- For every $x \in \mathcal{X}$, the mapping $K(\cdot, x) : \mathcal{G} \to [0, 1]$ is a probability measure on $(\mathcal{Y}, \mathcal{G})$,
- For every $E \in \mathcal{G}$, the mapping $K(E, \cdot) : \mathcal{X} \to [0, 1]$ is $(\mathcal{F}, \mathbb{B}_{[0,1]})$-measurable.

Remark 63. The Markov kernel can be interpreted as a random mapping form $\mathcal{X}$ to a probability measure on $(\mathcal{Y}, \mathcal{G})$. Hence, we will often denote it with $P_{Y|X}$ — that is, it defines a probability distribution of Y for every realization of X. For the special case of finite $\mathcal{X}$ and $\mathcal{Y}$, the Markov kernel $K \equiv P_{Y|X}$ is simply the transition probability matrix of size $|\mathcal{X}| \times |\mathcal{Y}|$.

Fact 64 (Disintegration Theorem)**.** *Suppose P_{XY} is a joint distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{Y}$ is standard Borel (complete separable metric space with the Borel sigma-algebra). Then, there exists a Markov kernel K, such that for any measurable $E \subset \mathcal{X} \times \mathcal{Y}$, we have*

$$P_{XY}(E) = \int_{\mathcal{X}} P_X(dx) K(E^x|x), \quad \textit{for } E^x := \{y \in \mathcal{Y} : (x, y) \in E_x\}.$$

We can now define the general version of conditional relative entropy, using the Markov kernel.

Definition 65. Suppose X is an $\mathcal{X}$-valued, and Y is a $\mathcal{Y}$-valued random variable. Let P_{XY} and Q_{XY} denote two joint distributions of (X, Y). Then, if $\mathcal{Y}$ is standard Borel (or 'nice' in the language used by Durrett), then the conditional relative entropy between $P_{Y|X}$ and $Q_{Y|X}$ given P_X is defined as

$$D_{\mathrm{kl}}(P_{Y|X} \parallel Q_{Y|X} | P_X) = \mathbb{E}_{P_X} \left[D_{\mathrm{kl}} \left(P_{Y|X}(\cdot, X) \parallel Q_{Y|X}(\cdot, X) \right) \right].$$

Remark 66. Having defined relative entropy, and conditional relative entropy for general distributions, we can immediately use them to define mutual information and conditional mutual information.

Remark 67. All the properties that we proved for the relative entropy of discrete or continuous random variables are also valid for the general setting considered here. These properties include the nonnegativity, convexity, and the DPI.

The main summary of this section is that we can define the relative entropy can be defined for distributions on very general observation spaces (such as on manifolds, or on function spaces). In particular, the differential entropy of a random variable X can be recovered as the relative entropy of P_X and an appropriate invariant measure, such as the Lebesgue measure for $\mathcal{X} = \mathbb{R}^d$.

2.2.1 Variational Definition I: Gelfand-Yaglom-Peres

In this section, and the next, we present two *variational definitions* of relative entropy for general probability distributions — that is, we define relative entropy as the solution of an optimization problem. There are several benefits of such a representation:

- several properties, such as convexity and lower semi-continuity, can be easily inferred,
- such representations easily allow us to get upper or lower bounds by considering specific values of the objective function being optimized.

Example 68. As a simple example, consider the ℓ_1 norm of a vector: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^d |x_i|$. It is easy to check that it can also be defined as follows:

$$\|\boldsymbol{x}\|_1 = \sup_{y \in \mathbb{R}^d : \|y\|_\infty \leq 1} \sum_{i=1}^d x_i y_i.$$

The above definition immediately implies that the map $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_1$ is convex: since it is the supremum of linear functions. Furthermore, we also immediately observe that it is lower semi-continuous: since it is the supremum of continuous functions. Furthermore, by choosing any specific value of $\boldsymbol{y}$, we get a lower bound on the ℓ_1 norm of $\boldsymbol{x}$.

We now present the first variational definition of relative entropy, which says that for general probability spaces, the relative entropy between two distributions is equal to the supremum of the relative entropy between quantized versions of the two distributions. In other words, for most purposes, analyzing the properties of relative entropy (and derived quantities, such as mutual information) for discrete distributions with finite support is without loss of generality.

Theorem 69. *Let $(\mathcal{X}, \mathcal{F})$ denote a measurable space, and let P and Q denote two probability measures on this space. Let $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$ denote a finite disjoint partition of $\mathcal{X}$, consisting of elements of $\mathcal{F}$. Then, we have*

$$D_{kl}(P, Q) = \sup_{\mathcal{E}} \sum_{E \in \mathcal{E}} P(E) \log\left(\frac{P(E)}{Q(E)}\right) = \sup_{\mathcal{E}} D_{kl}\left(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}\right),$$

where we have used $P_{\mathcal{E}}$ and $Q_{\mathcal{E}}$ to denote the quantized versions of P and Q over the partition $\mathcal{E}$.

Proof. We will proof the equality by showing that both $\leq$ and $\geq$ simultaneously hold. The lower bound is an easy consequence of the DPI for relative entropy. In particular, let $\mathcal{E}$ denote a partition of $\mathcal{X}$ with m elements $\{E_1, \ldots, E_m\}$. Let $f : \mathcal{X} \to [m]$ be a function defined as $f(x) = \sum_{i=1}^m i \mathbf{1}_{x \in E_i}$, and let $Y = f(X)$. Then, we have

$$D_{\mathrm{kl}}(P \parallel Q) \overset{(i)}{\geq} D_{\mathrm{kl}}\left(P_Y \parallel Q_Y\right) = D_{\mathrm{kl}}\left(P_{X,\mathcal{E}} \parallel Q_{X,\mathcal{E}}\right),$$

where (i) follows from the DPI for relative entropy. Since $\mathcal{E}$ above was arbitrary, we can take a supremum over all such finite partitions to get the lower bound.

Proving the other direction is more involved, and we proceed in the following steps:

Step 1: We begin by noting that without loss of generality, we can assume $P \ll Q$. Because, if $P \not\ll Q$, then both sides of the required equality are infinite. In particular, if $P \not\ll Q$, then there exists a measurable set E, such that $Q(E) = 0$ but $P(E) > 0$. Then, define $\mathcal{E} = \{E, E^c\}$, and observe that $D_{\mathrm{kl}}(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}) = \infty$.

Step 2: Since $P \ll Q$, there exists a measurable Radon-Nikodym derivative $X = dP/dQ$. Since $D_{\mathrm{kl}}(P \parallel Q) = \mathbb{E}_Q[\varphi(X)]$, for real valued sequence $c_n \to \infty$, we have $D_{\mathrm{kl}}(P \parallel Q) = \lim_{c \to \infty} \mathbb{E}_Q[\varphi(X) \mathbf{1}_{X \leq c}]$ by the monotone convergence theorem (MCT).

Step 3: Fix a $c > 0$ and $\epsilon > 0$, and let the integer n denote c/ϵ. Construct a partition $\mathcal{E} = \{E_0, \dots, E_n\}$, where the sets E_0 through E_n are defined as

$$\begin{aligned} & E_j = \{\omega : j\epsilon \leq X(\omega) \leq (j+1)\epsilon\}, \text{ for } j = 0, \dots, n-1, \\ \text{and} \quad & E_n = \{\omega : X(\omega) \geq c\}. \end{aligned}$$

Using this partition, define a discrete approximation of X, as

$$Y_n = \sum_{j=0}^{n-1} j\epsilon \mathbf{1}_{E_j}.$$

Let $X_c = X\mathbf{1}_{X \leq c}$, and note that $Y_n \leq X_c$ and $|Y_n - X_c| \leq \epsilon$. Now, note that for a fixed $c < \infty$, the function φ restricted to the domain $[0, c]$ is uniformly continuous. Hence, there exists a $\delta \equiv \delta(c, \epsilon)$, such that we have $|\varphi(Y_n) - \varphi(X_c)| \leq \delta$ almost surely. Furthermore, for a fixed c, the term δ goes to 0 as $\epsilon \to 0$.

The above uniform continuity result implies that

$$\mathbb{E}_Q[\varphi(X_c)] - \delta \leq \mathbb{E}_Q[\varphi(Y_n)] = \sum_{j=0}^{n-1} Q(E_j)\varphi(j\epsilon). \tag{2.6}$$

Now, we observe that

$$\mathbb{E}_Q[j\epsilon] \leq \mathbb{E}_Q[X\mathbf{1}_{E_j}] = P(E_j) \leq \mathbb{E}_Q[(j+1)\epsilon],$$

which implies that

$$j\epsilon \leq \frac{P(E_j)}{Q(E_j)} \leq (j+1)\epsilon \quad \text{or} \quad \left|\varphi\left(\frac{P(E_j)}{Q(E_j)}\right) - \varphi(j\epsilon)\right| \leq \delta.$$

Plugging this back into (2.6), we get

$$\mathbb{E}_Q[\varphi(X_c)] - \delta \leq \mathbb{E}_Q[\varphi(Y_n)] = \sum_{j=0}^{n-1} Q(E_j)\varphi\left(\frac{P(E_j)}{Q(E_j)}\right) + \delta = D_{\text{kl}}\left(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}\right) + \delta.$$

To complete the proof, we take $\epsilon \to 0$ for a fixed c, and then take $c \to \infty$. □

We record an immediate consequence of the above result.

Corollary 70. *For any $\epsilon > 0$, there exists a finite partition of $\mathcal{X}$, such that*

$$D_{kl}(P, Q) - \epsilon \leq D_{kl}(P_{\mathcal{E}}, Q_{\mathcal{E}}).$$

2.2.2 Variational Definition II: Donsker-Varadhan

We now present a functional variational representation of relative entropy, due to Donsker and Varadhan.

Theorem 71. *Consider a measurable space $(\mathcal{X}, \mathcal{F})$, with two probability measures P and Q. Suppose $P \ll Q$, and let $\mathcal{C}_Q$ denote the set of measurable functions $f : \mathcal{X} \to \mathbb{R}$, such that $\mathbb{E}_Q[e^f(X)] < \infty$. Then, we have*

$$D_{kl}(P \parallel Q) = \sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log\left(\mathbb{E}_Q[e^{f(X)}]\right).$$

Proof. First, note that we can restrict our attention to $P \ll Q$. For otherwise, if $P \not\ll Q$, then there must exist an E such that $Q(E) = 0$, but $P(E) > 0$. Then, define a function $f_c = c\mathbf{1}_E$, and note that it lies in $\mathcal{C}_Q$ for all values of $c \in \mathbb{R}$. Then, we have

$$\sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log E_Q[e^{f(X)}] \geq \sup_{c > 0} \mathbb{E}_P[f_c(X)] - \log E_Q[e^{f_c(X)}] = \sup_{c > 0} cP(E) = \infty.$$

Hence, we can assume that $P \ll Q$ for the rest of the proof. As in the previous case, we prove the result in two steps to show that both the $\leq$ and $\geq$ hold. For any $f \in \mathcal{C}_Q$, define the *tilted* distribution Q^f, such that for any measurable E, we have

$$Q^f(E) = \mathbb{E}_Q\left[e^{f(X)-Z_f}\mathbf{1}_E(X)\right], \quad \text{where} \quad Z_f = \log\left(\mathbb{E}_Q\left[e^{f(X)}\right]\right).$$

Now, observe that $Q^f \ll Q$, and

$$\log\left(\frac{dQ^f}{dQ}\right) = f(X) - Z_f,$$

which implies that

$$\mathbb{E}_P\left[\log\left(\frac{dQ^f}{dQ}\right)\right] = \mathbb{E}_P[f(X)] - Z_f.$$

Next, we note that we also have $Q \ll Q^f$. To see why this is true, suppose there exists an E such that $Q^f(E) = 0$, but $Q(E) > 0$. Define $A_0 = \{e^{f(X)-Z_f} \geq 1\} \cap E$, and for $i \geq 1$, define $A_i = \{e^{f(X)-Z_f} \in (1/(i+1), 1/i]\} \cap E$. Then, we have $A_i \cap A_j = \emptyset$ for $i \neq j$, and furthermore, $E = \cup_{i=0}^{\infty} A_i$. Due to the countable additivity, we have $Q(E) = \sum_{i=0}^{\infty} Q(A_i)$, which means that there must exist an n, such that $Q(A_n) > 0$. This implies that

$$0 = Q^f(E) = \sum_{i=0}^{\infty} Q^f(A_i) \geq \sum_{i=0}^{\infty} \frac{Q(A_i)}{i+1} \geq \frac{Q(A_n)}{n+1} > 0,$$

which is a contradiction. Hence, we have proved that $Q \ll Q^f$, which in turn implies that $P \ll Q^f$, and hence dP/dQ^f is well defined

Now, let G denote the set $\{dP/dQ > 0\}$ and note that $\mathbb{E}_P[\log(dQ^f/dQ)] = \mathbb{E}_P[\log(dQ^f/dQ)\mathbf{1}_G]$. Introduce the probability measures $\widetilde{Q}$ and $\widetilde{Q}^f$, such that $\widetilde{Q}(E) = Q(E \cap G)$, and $\widetilde{Q}^f(E) = \widetilde{Q}(E \cap G)$. Then, we can write

$$\begin{aligned}
\mathbb{E}_P\left[\log\left(\frac{dQ^f}{dQ}\right)\right] &= \mathbb{E}_P\left[\log\left(\frac{dQ^f}{dQ}\right)\mathbf{1}_G\right] = \mathbb{E}_P\left[\log\left(\frac{d\widetilde{Q}^f}{dP}\frac{dP}{d\widetilde{Q}}\right)\mathbf{1}_G\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{dP}{d\widetilde{Q}}\right)\mathbf{1}_G\right] - \mathbb{E}_P\left[\log\left(\frac{dP}{d\widetilde{Q}^f}\right)\mathbf{1}_G\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] - \mathbb{E}_P\left[\log\left(\frac{dP}{dQ^f}\right)\right] \\
&= D_{\mathrm{kl}}(P \parallel Q) - D_{\mathrm{kl}}(P \parallel Q^f) \leq D_{\mathrm{kl}}(P \parallel Q).
\end{aligned}$$

Since f was an arbitrary element of $\mathcal{C}_Q$, this completes the proof of one side of the inequality.

To show the other direction, we will rely on Theorem 69. Now consider the case when $P \ll Q$. Then, consider any partition $\mathcal{E}$ of the domain, and define f as

$$f = \sum_{E \in \mathcal{E}} \log\left(\frac{P(E)}{Q(E)}\right)\mathbf{1}_E.$$

For this function, the objective function is

$$\begin{aligned}
\mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^{f(X)}] &= \sum_E P(E)\log\left(\frac{P(E)}{Q(E)}\right) - \log\left(\sum_{E \in E} Q(E)\frac{P(E)}{Q(E)}\right) \\
&= D_{\mathrm{kl}}\left(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}\right)
\end{aligned}$$

For all choices of the partition $\mathcal{E}$, the corresponding f is still a simple function, and hence, it also belongs to $\mathcal{C}_Q$. Denoting the class of simple functions by $\mathcal{S}$, we then obtain

$$\sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log E_Q[e^{f(X)}] \geq \sup_{f \in \mathcal{S}} \mathbb{E}_P[f(X)] - \log E_Q[e^{f(X)}] \geq \sup_{\mathcal{E}} D_{\mathrm{kl}}(P_{\mathcal{E}}, Q_{\mathcal{E}}).$$

The last term is precisely the definition of $D_{\mathrm{kl}}(P \parallel Q)$ from Theorem 69. □

Remark 72. A simple consequence of the above result is that if $D_{\mathrm{kl}}(P \parallel Q)$ is small, then we should expect $\mathbb{E}_P[f(X)]$ to be well approximated by $\mathbb{E}_Q[f(X)]$. To see why, note that $e^x \approx 1 + x$ and $\log(1 + x) \approx x$ for small enough x. Thus choosing any function $f \in \mathcal{C}_Q$, we have

$$D_{\mathrm{kl}}(P \parallel Q) \geq \mathbb{E}_P[f(X)] - \log\left(\mathbb{E}_Q[e^{f(X)}]\right) \approx \mathbb{E}_P[f(X)] - \log\left(1 + \mathbb{E}_Q[f(X)]\right) \approx \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)].$$

Thus, a small value of $D_{\mathrm{kl}}(P \parallel Q)$ implies that the expected values of $f(X)$ under P and Q must be close to each other.

Since mutual information between (X, Y) is the relative entropy between the joint distribution and the product of marginals, we have the following corollary.

Corollary 73. *For $(X, Y) \sim P_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, let $\mathcal{C}$ denote the class of functions $\{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} : e^f \in L^1(P_X \times P_Y)\}$. Then, we have*

$$I(X; Y) = \sup_{f \in \mathcal{C}} \mathbb{E}_{P_{XY}}[f(X, Y)] - \log \mathbb{E}_{P_X \times P_Y}\left[e^{f(X,Y)}\right].$$

Application: Generalization bound in machine learning. In statistical learning theory, we are usually given a training dataset $S = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ consisting of n i.i.d. training points drawn from a distribution P. A learning algorithm, $\mathcal{A}$, is a channel (or a Markov kernel) from $\mathcal{Z}^n$ to a 'hypothesis class' $\mathcal{H}$. Given a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, we can define the population and empirical risk of a hypothesis $h \in \mathcal{H}$ as

$$L_P(h) = \mathbb{E}_{Z \sim P}[\ell(h, Z)], \quad \text{and} \quad L_S(h) = \frac{1}{n} \sum_{Z_i \in S} \ell(h, Z_i).$$

Now, let H denote the possibly random output of a learning algorithm $\mathcal{A}$ on a training set S. Then, the generalization error of this algorithm is the expected difference between the test and training risk of H:

$$\mathrm{gen}(P, P_{H|S}) = \mathbb{E}\left[L_P(H) - L_S(H)\right] = \mathbb{E}\left[L_{S'}(H) - L_S(H)\right],$$

where S' is an independent copy of S. Now, we can state the following bound over the generalization error.

Proposition 74 (Xu and Raginsky, 2016)**.** *Suppose the loss function is such that $\ell(h, Z)$ is σ^2-subGaussian for all $h \in \mathcal{H}$. Then, we have*

$$gen(P, P_{H|S}) \leq \sqrt{\frac{2\sigma^2 I(S; H)}{n}}.$$

Proof. To prove this result, we note the following:

$$I(S; H) = \sup_f \mathbb{E}_{P_{SH}}[f(S, H)] - \log \mathbb{E}_{P_S \times P_H}\left[e^{f(S,H)}\right]. \tag{2.7}$$

To get the bound, we will select a specific $f(s, h) = \frac{1}{n}\sum_{z_i \in s} \ell(h, z_i)$, and note that f is (σ^2/n)-subGaussian. Furthermore, note that the generalization error is equal to

$$\mathrm{gen}(P, P_{H|S}) = \mathbb{E}_{P_{SH}}\left[f(S, H)\right] - \mathbb{E}_{P_S \times P_H}[f(S', H')].$$

For any $\lambda \in \mathbb{R}$, plugging λf in (2.7), we get

$$\begin{aligned}
I(S; H) &\geq \mathbb{E}_{P_{SH}}\left[\lambda f(S, H)\right] - \log\left(\mathbb{E}_{P_S \times P_H}\left[e^{\lambda f(S', H')}\right]\right) \\
&\geq \mathbb{E}_{P_{SH}}\left[\lambda f(S, H)\right] - \log\left(\exp\left(\frac{\lambda^2\sigma^2}{2n} + \lambda \mathbb{E}_{P_S \times P_H}[f(S', H')\right)\right) \\
&= \mathbb{E}_{P_{SH}}[\lambda f(S, H)] - \mathbb{E}_{P_S \times P_H}[\lambda f(S', H')] - \frac{\lambda^2\sigma^2}{2n} \\
&= \lambda\, \mathrm{gen}(P, P_{H|S}) - \frac{\lambda^2\sigma^2}{2n}.
\end{aligned}$$

On optimizing for λ, we get

$$I(S;H) \geq \sup_{\lambda} \lambda \operatorname{gen}(P, P_{H|S}) - \frac{\lambda^2\sigma^2}{2n} = \frac{n \operatorname{gen}^2(P, P_{H|S})}{2\sigma^2}.$$

On rearranging, we get the required

$$\operatorname{gen}(P, P_{H|S}) \leq \sqrt{\frac{2\sigma^2 I(S;H)}{n}}.$$

□

Now, suppose h^* denotes the element of $\mathcal{H}$ that achieves the minimum expected loss. How does $\mathbb{E}[L_P(H)]$ compare with $L_P(h^*)$? To analyze this, note that

$$\mathbb{E}[L_P(H)] \leq \operatorname{gen}(P, P_{H|S}) + \mathbb{E}_{P_{SH}}\left[L_S(H)\right].$$

Hence, on subtracting $L_S(h^*)$ on both sides, and using Proposition 74, we get

$$\mathbb{E}[L_P(H)] - L_P(h^*) \leq \sqrt{\frac{2\sigma^2 I(S;H)}{n}} + \mathbb{E}_{P_{SH}}\left[L_S(H) - L_S(h^*)\right].$$

Consider the following two cases:

- If $|\mathcal{H}| < \infty$, then we can upper bound $I(S;H)$ with $H(H) \leq \log(|\mathcal{H}|)$.
- More generally, suppose we consider an appropriate ϵ-covering set of $\mathcal{H}$, denoted by $|\mathcal{H}_\epsilon|$. Then, we can bound $I(S;H)$ with $\log(|\mathcal{H}_\epsilon|)$, which recovers the metric entropy bound.

Note that if H is selected by an empirical risk minimization (ERM) method, the second term on the RHS is upper bounded by 0.

2.3 f-divergences

Definition 75. Suppose $f : (0,\infty) \to \mathbb{R}$ is a convex function with $f(1) = 0$, and $f(0) \coloneqq \lim_{x\to 0} f(x)$. Then, the f-divergence between two distributions P and Q, with $P \ll Q$ is defined as

$$D_f(P \parallel Q) \coloneqq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}(X)\right)\right], \quad \text{if } P \ll Q.$$

For general P and Q, let μ denote a common dominating measure (such as $P+Q$), and $q = dQ/d\mu$ and $p = dP/d\mu$. Then, the f-divergence between P and Q is defined as

$$D_f(P \parallel Q) = \int_{q>0} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu + \mathbb{P}\left(q = 0\right) \lim_{x\to\infty} x f(1/x).$$

If $\mathbb{P}(q=0) = 0$, then the second term is assumed to be 0, irrespective of $\lim_{x\to\infty} x f(1/x)$.

Remark 76. For discrete distributions, we have the following simple definition of f-divergences:

$$D_f(P \parallel Q) = \sum_{x\in\mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

Remark 77. The above definition unifies several popular statistical divergences, such as

- Relative entropy, with $f(x) = x\log x$
- Total variation, with $f(x) = \frac{1}{2}|x-1|$

- Squared Hellinger distance, with $f(x) = (1-\sqrt{x})^2$
- χ^2-distance, with $f(x) = (x-1)^2$.

We can develop several properties of f-divergences, that parallel those for relative entropy: see Polyanskiy and Wu [2023, Chapter 7] for a thorough discussion. Below, we state and prove the nonnegativity, convexity, and DPI for f-divergences.

Proposition 78. *For any convex $f : [0, \infty]$, with $f(1) = 0$, and $f(0) := \lim_{x \downarrow 0} f(x)$, we have*

(a) $D_f(P \parallel Q) \geq 0$, for all P, Q. Furthermore, if f is strictly convex, then the equality holds only for $P = Q$.

(b) $D_f(P_{XY} \parallel Q_{XY}) \geq D_f(P_X \parallel Q_X)$.

(c) The mapping $(P, Q) \mapsto D_f(P \parallel Q)$ is convex.

(d) Suppose $P_{XY} = P_X \times P_{Y|X}$ and $Q_{XY} = Q_X \times P_{Y|X}$. Then, we have the following data-processing inequality:

$$D_f(P_Y \parallel Q_Y) \leq D_f(P_X \parallel Q_X).$$

Proof. We will prove the statements under the simplifying assumption that $P_{XY} \ll Q_{XY}$.

(a) The result follows from a direct application of Jensen's inequality. In particular, we have

$$D_f(P \parallel Q) = \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}(X) \right) \right] \geq f\left(\mathbb{E}_Q \left[\frac{dP}{dQ}(X) \right] \right) = f(1) = 0.$$

If f is not affine, then the equality holds only if dP/dQ is a constant Q-almost-surely.

(b) For simplicity, we will prove this monotonicity result for the simple case of discrete distributions.

$$\begin{aligned} D_f(P_{XY} \parallel Q_{XY}) &= \sum_{x,y} q(x,y) f\left(\frac{p(x,y)}{q(x,y)} \right) = \sum_{x \in \mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} q(y|x) f\left(\frac{p(x)p(y|x)}{q(x)q(y|x)} \right) \\ &\geq \sum_{x \in \mathcal{X}} q(x) f\left(\sum_{y \in \mathcal{Y}} q(y|x) \frac{p(x)p(y|x)}{q(x)q(y|x)} \right) \\ &\geq \sum_{x \in \mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)} \sum_{y \in \mathcal{Y}} p(y|x) \right) = \sum_{x \in \mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)} \right) = D_f(P_X \parallel Q_X). \end{aligned}$$

(c) The proof of this result uses the monotonicity. In particular, let P_0, P_1 and Q_0, Q_1 denote two pairs of distributions. Now, let $X \sim \text{Bernoulli}(\lambda)$ for some $\lambda \in [0,1]$. Then, define the following joint distributions (P_{XY} and Q_{XY}):

$$P_{Y|X=0} = P_0, \quad P_{Y|X=1} = P_1, \quad \text{and} \quad Q_{Y|X=0} = Q_0, \quad Q_{Y|X=1} = Q_1.$$

Then, by the monotonicity result, we have

$$\begin{aligned} D_f(P_{XY} \parallel Q_{XY}) &= \overline{\lambda} \sum_y Q_0(y) f\left(\frac{\overline{\lambda} P_0(y)}{\overline{\lambda} P_1(y)} \right) + \lambda \sum_y Q_1(y) f\left(\frac{\lambda P_1(y)}{\lambda Q_1(y)} \right) \\ &\geq D_f(P_Y \parallel Q_Y) \\ &= D_f\left(\overline{\lambda} P_0 + \lambda P_1 \parallel \overline{\lambda} Q_0 + \lambda Q_1 \right). \end{aligned}$$

(d) Again, we prove this in the simple setting of discrete distribution. By monotonicity, we know that

$$D_f(P_Y \parallel Q_Y) \leq D_f(P_{XY} \parallel Q_{XY}).$$

Note that since the conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ are the same (i.e., the marginals P_X and P_Y are passed through the same channel), we have

$$\begin{aligned} D_f(P_{XY} \parallel Q_{XY}) &= \sum_{x \in \mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} q(y|x) f\left(\frac{p(x)p(y|x)}{q(x)q(y|x)}\right) \\ &= \sum_{x \in \mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} q(y|x) f\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_{x \in \mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \sum_{y \in \mathcal{Y}} q(y|x) = D_f(P_X \parallel Q_X). \end{aligned}$$

□

As for the case of relative entropy, we can obtain the analogs of the two variational definitions for f-divergences as well.

Theorem 79. *Let P and Q be any two probability measures defined on some measurable space $(\mathcal{X}, \mathcal{F})$. For any finite disjoint partition of $\mathcal{X}$, denoted by $\mathcal{E} = \{E_i \in \mathcal{F} : 1 \leq i \leq n\}$, let $P_{\mathcal{E}}$ and $Q_{\mathcal{E}}$ denote the discrete distributions with pmfs $\{P(E) : E \in \mathcal{E}\}$ and $\{Q(E) : E \in \mathcal{E}\}$. Then, we have the following:*

$$D_f(P \parallel Q) = \sup_{\mathcal{E}} D_f(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}) = \sup_{\mathcal{E}} \sum_{E \in \mathcal{E}} Q(E) f\left(\frac{P(E)}{Q(E)}\right).$$

The proof of this statement follows the general argument used in proving Theorem 69 for the analogous result for relative entropy. We refer the reader to Polyanskiy and Wu [2023, § 7.14] for details.

Next, we state and prove an analog of Theorem 71 for f-divergences on $\mathcal{X} = \mathbb{R}^d$.

Theorem 80. *Let $P \ll Q$ be two distributions on $\mathcal{X} = \mathbb{R}^d$, and furthermore assume that both P and Q admit densities (p, and q respectively) w.r.t. the Lebesgue measure μ. With f^* denoting the convex-conjugate of f, we have the following variational representation of $D_f(P \parallel Q)$:*

$$D_f(P \parallel Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}(X)\right)\right] = \sup_{g: \mathcal{X} \to \mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[f^*(g(X))\right],$$

where g is restricted to ensure that both expectations are finite.

Proof. The proof follows directly from the definition of convex conjugates. First note that since f is convex, we have $f = (f^*)^*$. Hence, for any $u \in (0, \infty)$, we have

$$f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v).$$

Thus, with $u = p(x)/q(x)$, we have

$$f\left(\frac{p(x)}{q(x)}\right) = \sup_{v \in \mathbb{R}} \frac{p(x)v}{q(x)} - f^*(v).$$

For any function g, plugging the value of $v = g(x)$ above gives us a lower bound on $f(p(x)/q(x))$. Hence, for an arbitrary function g, we have

$$D_f(P \parallel Q) \geq \sup_g \int_{\mathcal{X}} q(x) \left(g(x)\frac{p(x)}{q(x)} - f^*(g(x))\right) dx = \sup_g \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))].$$

To show the other direction: fix an arbitrary $\epsilon > 0$, and define g_ϵ as the function with values $g(x) = v_x$; where v_x ensures $v_x p(x)/q(x) - f^*(v_x) \geq f(p(x)/q(x)) - \epsilon$. Plugging this in, we get

$$D_f(P \parallel Q) \leq \mathbb{E}_P[g_\epsilon(X)] - \mathbb{E}_Q[f^*(g_\epsilon(X))] + \epsilon.$$

□

Example 81 (χ^2-divergence)**.** Recall that χ^2 distance corresponds to $f(x) = (x-1)^2$. The convex conjugate of this function is equal to $f^*(u) = u^2/4 + u = (u/2+1)^2 - 1$. To see this, note that

$$f^*(u) = \sup_{v \in \mathbb{R}} uv - (v-1)^2 = sup_{v\in\mathbb{R}} g(v)$$

Note that g is a concave function, and $g'(v) = u - 2(v-1)$. Hence, its optimal value is achieved at $v^* = 1 + u/2$, and is equal to

$$f^*(u) = u\left(\frac{u}{2}+1\right) - \frac{u^2}{4} = \frac{u^2}{4} + u = \left(\frac{u}{2}+1\right)^2 - 1.$$

Plugging this into the variational definition of f-divergences, we get

$$\begin{aligned}\chi^2(P \parallel Q) &= \sup_g \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[\left(\frac{g(X)}{2}+1\right)^2\right] + 1 \\ &= \sup_g 2\mathbb{E}_P\left[\left(\frac{g(X)}{2}+1\right)\right] - \mathbb{E}_Q\left[\left(\frac{g(X)}{2}+1\right)^2\right] - 1 \\ &= \sup_h 2\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h^2(X)] - 1,\end{aligned}$$

where in the last inequality, we used the change of variable $h(x) \leftarrow g(x)/2 + 1$.

Note that the above variational definition is not scale-invariant, and we can replace $h(\cdot)$ with an arbitrary $\lambda h(\cdot)$, to get

$$\chi^2(P \parallel Q) = \sup_{h,\lambda} 2\lambda \mathbb{E}_P[h(X)] - \lambda^2 \mathbb{E}_Q[h^2(X)] - 1.$$

On optimizing over λ, we get the following, more interpretable, definition

$$\chi^2(P \parallel Q) = \sup_h \frac{\left(\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)]\right)^2}{\mathbb{V}_Q\left(h(X)\right)}.$$

That is, two distributions P and Q are distinguishable in χ^2-divergence, if there exists a function h, for which the the absolute difference in means under P and Q is much larger than the standard deviation of $h(X)$ under Q.

Application**: Estimation error bounds.** Let us consider an estimation problem in the simplest case, where the parameter set Θ is a subset of $\mathbb{R}$. Now, let θ denote the true parameter, $X \sim P_\theta$, and $\widehat{\theta} = \widehat{\theta}(X)$ is some estimate of θ based on X. That is, we have the Markov chain $\theta \to X \to \widehat{\theta}$.

Proposition 82 (Hammersley-Chapman-Robbins)**.** *In the setting described above, the expected squared error of any estimator $\widehat{\theta}$ satisfies*

$$\mathbb{E}_\theta[(\theta - \widehat{\theta})^2] \geq \mathbb{V}_\theta(\widehat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{\left(\mathbb{E}_\theta[\widehat{\theta}(X)] - \mathbb{E}_{\theta'}[\widehat{\theta}(X)]\right)^2}{\chi^2\left(P_{\theta'} \parallel P_\theta\right)}.$$

Proof. This result follows by a simple application of the DPI for f-divergences, and the variational definition of χ^2-divergence obtained in Example 81. In particular, let $\theta' \in \Theta$ such that $\theta' \neq \theta$. Then, with $P_X = P_{\theta'}$ and $Q_X = P_\theta$, we have

$$\chi^2\left(P_X \parallel Q_X\right) \overset{(i)}{\geq} \chi^2\left(P_{\widehat{\theta}} \parallel Q_{\widehat{\theta}}\right) \overset{(ii)}{\geq} \frac{\left(\mathbb{E}_\theta[\widehat{\theta}] - \mathbb{E}_{\theta'}[\widehat{\theta}]\right)^2}{\mathbb{V}_\theta(\widehat{\theta})},$$

where (i) is due to the data-processing inequality, and (ii) is due to the variational representation of χ^2-divergence with $h(x) = x$. On rearranging, we get

$$\mathbb{V}_\theta(\widehat{\theta}) \geq \frac{\left(\mathbb{E}_\theta[\widehat{\theta}] - \mathbb{E}_{\theta'}[\widehat{\theta}]\right)^2}{\chi^2(P_{\theta'} \parallel P_\theta)}.$$

As $\theta' \neq \theta$ is an arbitrary element of Θ, the result follows by taking a supremum over all such θ'. □

Corollary 83 (Cramer-Rao lower bound). *Suppose that the distributions $\{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ admit densities (w.r.t. Lebesgue measure) that are twice continuously differentiable. If we restrict our attention to unbiased $\widehat{\theta}$, then Proposition 82 implies the following:*

$$\mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \mathbb{V}_\theta(\widehat{\theta}) \geq \frac{1}{J_F(\theta)}, \quad \textit{where} \quad J_F(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log(p_\theta(X))}{\partial\theta}\right)^2\right].$$

Proof. The conclusion follows from a fact about χ^2-distance that

$$\chi^2(P_{\theta'} \parallel P_\theta) = J_F(\theta)(\theta - \theta')^2 + o\left((\theta - \theta')^2\right).$$

To see why this is true, consider the definition of χ^2-divergence between $P_{\theta'}$ and $P_{\theta'}$

$$\chi^2(P_{\theta'} \parallel P_\theta) = \int_{\mathbb{R}} p_\theta(x)\left(\frac{p_{\theta'}(x)}{p_\theta(x)} - 1\right)^2 dx = \int_{\mathbb{R}} \frac{(p_{\theta'}(x) - p_\theta(x))^2}{p_\theta(x)} dx.$$

Now, by a Taylor' expansion around $p_\theta(x)$, we have

$$p_{\theta'}(x) = p_\theta(x) + \dot{p}_\theta(x)(\theta' - \theta) + o\left(|\theta' - \theta|\right).$$

Plugging this approximation back into the definition of $\chi^2(P_{\theta'} \parallel \theta)$, we get that

$$\chi^2(P_{\theta'} \parallel P_\theta) = (\theta' - \theta)^2 \int \frac{\dot{p}_\theta^2(x)}{p_\theta(x)} dx + o\left((\theta' - \theta)^2\right).$$

□

Chapter 3

Compression and Gambling

In this chapter, we introduce the problem of compression, or source coding, where the goal is to assign sequences (often binary) to represent symbols drawn from a known probabilistic source in the most efficient (i.e., with smallest average length) manner. We establish that entropy of a distribution characterizes the fundamental limit as the optimal compression rate; which can be achieved by coding schemes that assign shorter codewords to more probable symbols. A major theme in this chapter is the equivalence between compression, and probability assignment. This connection is formalized through Kraft-Macmillan's Lemma in one direction (from codewords to probability distributions), and through the Shannon-Fano-Elias coding scheme in the other direction (from probability to codewords). Interestingly, the probability assignment viewpoint also leads to surprising connections between compression and gambling on horse races, and we end the chapter with a discussion on this.

Throughout this chapter we will focus on distributions over a discrete alphabet with countable support, denoted by $\mathcal{X}$. The task of compression then reduces to mapping elements of $\mathcal{X}$ to sequences of another D-ary alphabet $\mathcal{D}$. Often, we will set $\mathcal{D} = \{0, 1\}$, and focus on binary source codes. Accordingly, most of the logarithms in this chapter will be to the base 2 (some might be to the base $D \geq 2$).

3.1 Source Codes

We begin with a formal definition of source codes.

Definition 84. A D-ary source code for symbols from an alphabet $\mathcal{X}$ is a mapping $C : \mathcal{X} \to \mathcal{D}^*$, where $\mathcal{D}^* = \cup_{m=1}^{\infty} \mathcal{D}^m$. For any x, we refer to $C(x)$ as the codeword associated with x, and use $l(x)$ to denote its length. The average codeword length of an $\mathcal{X}$-valued random variable $X \sim P_X$, with the coding scheme C, is defined as $L_C(X) = \sum_{x \in \mathcal{X}} p_X(x) l(x)$.

For any $k \geq 1$, the k-extension of C, is the mapping from $\mathcal{X}^k \to \mathcal{D}^*$ that assigns $x^k = (x_1, \ldots, x_k)$ to $C(x^k) = C(x_1)C(x_2)\ldots C(x_k)$.

Example 85. Suppose $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$, and $P_X = (1/2, 1/4, 1/8, 1/8)$. Consider the binary coding scheme that assigns the codewords:

$$C(x_1) = 0, \quad C(x_2) = 10, \quad C(x_3) = 110, \quad C(x_4) = 111.$$

The average codeword length, L_C, of this coding scheme is 1.75, which is also equal to the entropy of the distribution P_X.

Most practical applications involve encoding large blocks or streams of symbols (think of a text document as a large block of letters and punctuation marks). After obtaining a compressed representation, we want the ability to reconstruct (or decode) the original block of symbols in a lossless manner. This requirement imposes certain constraints on the coding schemes, and we formally describe the nested class of schemes (with increasingly stronger constraints).

Definition 86. We are interested mainly in the following three classes of codes:

- We say a code C is *nonsingular*, if $x \neq x' \implies C(x) \neq C(x')$.
- We say a code C is *uniquely decodable*, if for all $k \geq 1$, the k extension of C is nonsingular.
- We say a code C is *instantaneously decodable*, if no codeword $C(x)$ is the prefix for another codeword $C(x')$.

Example 87. Let the alphabet $\mathcal{X}$ be as in Example 85, and consider the four coding schemes described in Table 3.1. From the table, it is easy to see that

- The coding scheme C_1 is clearly singular, since it assigns the same codeword 0 to multiple symbols (x_1 and x_3).
- The scheme C_2 is non-singular as it is a one-to-one mapping from the symbols to codewords. However, it is not uniquely decodable: the sequences 100 could either be decode as the single symbol x_3, or the triple $x_2x_1x_1$.
- The coding scheme C_3 is non-singular, and in fact is uniquely decodable (UD). To see why, process any codeword sequence as follows:
  - If first two bits are either 00 or 10, immediately decode them as x_2 and x_1 respectively, and continue.
  - If the first two bits are 11, then check the value of the next bit. If this is 1, then decode the first two bits to x_3, and continue processing the sequence starting with the third bit.
  - If the first three bits are 110, then count the total number of zeros following the first 11. If the number of zeros is even (say $2k$), then the first symbol is x_3, followed by k repetitions of x_2. If the the number of zeros is odd (say $2k+1$), then the first symbol is x_4, followed by k repetitions of x_2.

  Thus, despite being uniquely decodable, we may have to process the entire sequence of bits, before being able to decode even the first symbol.
- Finally, the scheme C_4 is prefix-free; and thus is instantaneously decodable.

Symbol (x)	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$
x_1	0	0	10	0
x_2	1	1	00	10
x_3	0	100	11	110
x_4	01	101	110	111

Table 3.1: The table defines four source codes for symbols in the alphabet $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$. The first scheme C_1 is clearly singular; C_2 is non-singular but not uniquely decodable; C_3 is uniquely decodable but not prefix-free (or instantaneously decodable); while C_4 is an example of a prefix-free code.

Summary. It is impossible to encode (in a lossless manner) even one symbol with singular codes. Non-singular codes cannot be applied to streams of symbols without potentially losing information. Uniquely decodable (UD) codes may require waiting arbitrarily long before even one symbol can be decoded. Instantaneously decodable (or prefix) codes are self-punctuating, and hence they can be decoded on a per-symbol basis.

3.2 Kraft's Inequality

Following the discussion in the previous section, we would like to construct prefix-free codes with minimum expected lengths. Clearly, we cannot assign very short codewords to all symbols without violating the prefix-free property. In this section, we obtain a precise characterization of the set of codeword lengths that can be assigned to prefix-free (in fact UD too) codes for countable alphabets $\mathcal{X}$.

Theorem 88. *Suppose $\mathcal{X}$ is a countable alphabet, and C denotes any binary prefix code with lengths $(l_i)_{i\geq 1}$. Then, we have*

$$\sum_{i\geq 1} 2^{-l_i} \leq 1. \tag{3.1}$$

Conversely, for any $(l_i)_{i\geq 1}$ satisfying (3.1)*, we can construct a prefix code with these lengths.*

Remark 89. When the alphabet $\mathcal{X}$ is finite, then the above result has a visual representation. Using the codewords, we can construct a binary tree (for example, with 0 denoting the left child, and 1 denoting the right child). Thus, each codeword corresponds to the path from the root to a leaf of the tree so obtained. Since the code is prefix-free, no codeword is an ancestor of another. Then, in this case, Kraft's inequality is a statement of the fact that the sum of the number of descendants of l_i (including it) at the level $l_{\max} = \max_j l_j$ is $\sum_i 2^{l_{\max}-l_i}$: a number which cannot be larger than $2^{l_{\max}}$.

Proof. For any binary codeword $\boldsymbol{y}_i = (y_1, y_2, \ldots, y_{l_i}) \in \{0,1\}^*$, we introduce the following mapping:

$$V(\boldsymbol{y}) = 0.y_1y_2\ldots y_{l_i} = \sum_{j=1}^{l_i} y_j 2^{-j}.$$

In words, V assigns $\boldsymbol{y}_i$ to a real number in $[0,1]$, whose binary representation is given by $0.\boldsymbol{y}_i$. Now, let $I(\boldsymbol{y}_i)$ denote the interval that consists of all binary sequences whose first l_i bits are equal to $\boldsymbol{y}_i$. It is easy to verify that

$$I(\boldsymbol{y}_i) = \left[V(\boldsymbol{y}_i), V(\boldsymbol{y}_i) + 2^{-l_i}\right).$$

Next, we observe that since C is a prefix code, no codeword is a prefix for another. This implies that

$$I(\boldsymbol{y}_i) \cap I(\boldsymbol{y}_j) = \emptyset, \quad \text{for } i \neq j.$$

Since the length of each interval $I(\boldsymbol{y}_i)$ is equal to 2^{-l_i}, and they are all contained in $[0,1]$, the sum of their lengths must be upper bounded by 1, as required.

For the converse part, suppose we are given lengths $(l_i)_{i\geq 1}$ that are sorted in increasing order, and satisfy Kraft's inequality. We can construct codewords with these lengths, as the as the binary representation of the smallest l_i bit binary number:

$$I_i = \left[\sum_{j=1}^{i-1} 2^{-l_j}, \sum_{j=1}^{i} 2^{-l_j}\right), \quad \text{for all } i \geq 1.$$

The length of interval I_i is equal to 2^{-l_i}, which by assumption sums up to no larger than 1. □

Can we gain anything by considering uniquely decodable codes? The next result says no.

Theorem 90. *Suppose C is a uniquely decodable code over a countable alphabet $\mathcal{X}$, with lengths $(l_i)_{i\geq 1}$. Then, the lengths must satisfy Kraft's inequality.*

Furthermore, for any $(l_i)_{i\geq 1}$ satisfying Kraft's inequality, there exists an instantaneous (hence also UD) code over $\mathcal{X}$ with those lengths.

Proof. If C is a UD code for symbols from $\mathcal{X}$, then it is also a UD code for any finite subset $\mathcal{X}'$ of $\mathcal{X}$. Let $\mathcal{X}_N$ denote the subset of $\mathcal{X}$ consisting of its first N elements. Then, note that

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} = \lim_{N\to\infty} \sum_{x\in\mathcal{X}_N} 2^{-l(x)}.$$

This means that it suffices to establish Kraft's inequality for an arbitrary finite N, since the bound follows by taking N to infinity.

Fix an $N < \infty$, and let $l_{\max} = \max_{x \in \mathcal{X}_N} l(x)$. Then, observe the following for an arbitrary $k \in \mathbb{N}$:

$$\begin{aligned}\left(\sum_{x\in\mathcal{X}_N} 2^{-l(x)}\right)^k &= \sum_{x_1\in\mathcal{X}_N}\sum_{x_2\in\mathcal{X}_N}\cdots\sum_{x_k\in\mathcal{X}_N} 2^{-l(x_1)}2^{-l(x_2)}\cdots 2^{-l(x_k)} \\ &= \sum_{x^k\in\mathcal{X}_N^k} 2^{-l(x^k)} \\ &= \sum_{m=1}^{kl_{\max}} a(m)2^{-m},\end{aligned}$$

where $a(m)$ denotes the number of sequences in $\mathcal{X}_N^k$ that are assigned a codeword of length $m \in \{1, 2, \ldots, kl_{\max}\}$. Next, we make the observation that $a(m) \leq 2^m$. This is because the code C is assumed to be uniquely decodable, which means that each codeword in $\{0,1\}^m$ is assigned to at most one element of $\mathcal{X}_N^k$. This implies that $a(m)2^{-m} \leq 1$, and thus

$$\left(\sum_{x\in\mathcal{X}_N} 2^{-l(x)}\right)^k \leq kl_{\max} \quad \Longrightarrow \sum_{x\in\mathcal{X}_N} 2^{-l(x)} \leq k^{1/k} l_{\max}^{1/k}.$$

The above inequality is true for all values of $k \geq 1$, and hence, by taking the limit of $k \to \infty$, we get

$$\sum_{x\in\mathcal{X}_N} 2^{-l(x)} \leq \lim_{k\to\infty} (kl_{\max})^{1/k} = \exp\left(\lim_{k\to\infty} \frac{1}{k}\ln(kl_{\max})\right) = 1.$$

Since N was arbitrary, we can then conclude that

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} = \lim_{N\to\infty} \sum_{x\in\mathcal{X}_N} 2^{-l(x)} \leq 1.$$

Thus, we have proved that any UD code must satisfy Kraft's inequality.

To show the converse, we simply observe that given $(l_i)_{i\in\mathbb{N}}$, we saw in Theorem 88 how to construct a prefix-free code with those lengths. Since prefix-free codes are also UD, the result follows. □

Remark 91. The non-overlapping intervals perspective used in proving Theorem 88 leads to an interesting conclusion regarding codes which satisfy Kraft's inequality strictly. That is if $\sum_{x\in\mathcal{X}} 2^{-l(x)} < 1$, that means the union of all the intervals is a strict subset of $[0, 1)$. That means there exists another nonempty interval in $[0, 1)$, not intersecting with any of the codeword intervals. Since every interval contains an irrational number, whose binary representation is non-terminating, we can conclude that there exist infinitely long sequences that cannot be decoded into symbols from $\mathcal{X}^*$ under such coding schemes.

3.3 Optimal Code Length, Practical Schemes, and Redundancy

Kraft's inequality gives us a precise mathematical characterization of the prefix-free and UD codes, which allows us to formulate the problem of finding optimal prefix-free codes as a constrained optimization problem. Formally, for a random variable X with distribution P over a countable alphabet $\mathcal{X}$, we can write the optimal coding problem as

$$L^*(X) \coloneqq \min_{l(x):x\in\mathcal{X}} \sum_{x\in\mathcal{X}} p(x)l(x), \quad \text{subject to } \sum_{x\in\mathcal{X}} 2^{-l(x)} \leq 1,\ l(x) \in \mathbb{N}. \tag{3.2}$$

Our first result shows that the optimal codeword length $L^*(X)$ is close to the entropy of X.

Theorem 92. *Suppose $X \sim P_X$ is a random variable taking values in a countable alphabet $\mathcal{X}$. Let $L^*(X)$ denote the average codeword length of the optimal uniquely decodable code, defined in (3.2). Then, we have the following:*

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

Proof. Upper bound. To show that $H(X)+1$ is an upper bound on $L^*(X)$, we simply need to show that there exists a UD source code, whose average length is no larger than $H(X)+1$. Define

$$l_x = \left\lceil \log\left(\frac{1}{p(x)}\right)\right\rceil \quad \text{for all } x \in \mathcal{X},$$

and observe that

$$\sum_{x\in\mathcal{X}} 2^{-l_x} \le \sum_{x\in\mathcal{X}} 2^{\log p(x)} = \sum_{x\in\mathcal{X}} p(x) = 1.$$

Thus, these collection of lengths satisfy Kraft's inequality. Hence, by the converse of Kraft's inequality, we know that there exists a prefix-free code, C, with these lengths. Furthermore, the average codeword length, $L_C(X)$, satisfies

$$L_C(X) = \sum_{x\in\mathcal{X}} p(x) l_x \le \sum_{x\in\mathcal{X}} p(x)\left(\log(1/p(x))+1\right) = H(X)+1.$$

This completes the proof of the upper bound.

Lower bound. We will show that the average length of any UD code C (i.e., any code that satisfies Kraft's inequality) must be loewr bounded by the entropy, which also implies the same about the optimal code. In particular, introduce the term $A = \sum_{x\in\mathcal{X}} 2^{-l(x)}$, and note that it is less than or equal to 1. Define the probability distribution R over $\mathcal{X}$, with p.m.f. $r(x) = 2^{-l(x)}/A$. Then, we have the following:

$$\begin{aligned}
L_C(X) - H(X) &= \sum_{x\in\mathcal{X}} p(x)l(x) - \sum_{x\in\mathcal{X}} p(x)\log(1/p(x)) \\
&= \sum_{x\in\mathcal{X}} p(X)\log\left(\frac{p(x)}{2^{-l(x)}}\right) \\
&= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(X)}{r(X)}\right) + \sum_{x\in\mathcal{X}} p(X)\log(1/A) \\
&= D_{\text{kl}}(P_X \parallel R) - \log(1/A) \ge D_{\text{kl}}(P_X \parallel R) \ge 0.
\end{aligned}$$

In the last inequality we used the nonnegativity of relative entropy, while in the second-last inequality, we used the fact that $A \le 1$ for uniquely decodable codes. □

Remark 93. The proof showing the lower bound in the above result also tells us the conditions under which $L^*(X) = H(X)$. It happens when (i) $A = 1$, and $P = R$. Or in other words, the equality is achieved when the distribution P is a *dyadic* distribution of the from $P(x) = 2^{-l_x}$ for all $x \in \mathcal{X}$.

Remark 94. Suppose X is a Bernoulli random variable with parameter p. Note that for small enough values of p, the entropy $h_2(p)$ can be made arbitrarily close to 0. However, the length of the optimal coding scheme is equal to 1. Thus the upper bound of Theorem 92 cannot be improved.

Remark 95. Theorem 92 states that the optimal codeword length can be up to one bit larger than the entropy. Nevertheless, we can do better by constructing codes for blocks of symbols at once, and spreading this additional bit over the entire block. That is, for some $n \ge 1$, we can construct an optimal code for the elements of the product space $\mathcal{X}^n$ (assuming computation is not an issue). By Theorem 92, we know that

$$H(X^n) \le L^*(X^n) \le H(X^n)+1, \quad \text{which implies} \quad H(X) \le \frac{1}{n}L^*(X^n) \le H(X) + \frac{1}{n}.$$

Or in other words, by considering large blocks of symbols at once, we can make the average codeword length per symbol arbitrarily close to the entropy.

Typical Set Coding. Before discussing the optimal Huffman coding scheme, we present a suboptimal, but conceptually simpler scheme. This builds upon the block coding idea, and works with symbols from $\mathcal{X}^n$ for large values of n. In particular, given X^n drawn i.i.d. from a distribution P, recall the definition of a typical set $A_n(\epsilon)$ as

$$A_n(\epsilon) = \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n}\sum_{i=1}^n p(x_i) - H(X) \right| \leq \epsilon \right\}.$$

For large enough n, we know that $A_n(\epsilon)$ satisfies the following properties:

- $\mathbb{P}(A_n(\epsilon)) \geq 1 - \epsilon$.
- $2^{n(H(X)-\epsilon)} \leq |A_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$.
- $2^{-n(H(X)+\epsilon)} \leq p^n(x^n) \leq 2^{-n(H(X)-\epsilon)}$, for all $x^n \in A_n(\epsilon)$.

Using these properties, we can define the following coding scheme:

- Enumerate all the elements of the typical set from 1 to $|A_n|$.
- Encode each sequence in the typical set by the binary code of its position in the enumeration (i.e., its position between 1 and $|A_n|$).
- Prefix the binary code of each typical sequence with a leading 0
- Enumerate all the sequences in the non-typical set; assign them the codeword of their binary representation of their position; prefix each codeword with an additional 1 to indicate it is a non-typical sequence.

Then, the average codeword length of this coding scheme, per symbol, satisfies

$$\begin{aligned}
\frac{L_C(X^n)}{n} &\leq (1-\epsilon)\frac{\log(|A_n|)+2}{n} + \epsilon\frac{n\log(|\mathcal{X}|)+2}{n} \\
&= \frac{\log(|A_n|)+2}{n} + \epsilon\frac{n\log(|\mathcal{X}|)-\log(|A_n|)}{n} \\
&\leq \frac{nH(X)+n\epsilon+2}{n} + \epsilon\frac{n\log(|\mathcal{X}|)-nH(X)+n\epsilon}{n} \\
&= H(X) + \epsilon\left(1+\log(|\mathcal{X}|)+\epsilon\right) + \frac{2}{n}.
\end{aligned}$$

Since $\epsilon > 0$ can be made arbitrarily small, by selecting large enough values of n, the typical set coding scheme achieves the optimal per-symbol compression rate asymptotically.

Huffman Coding. When $\mathcal{X}$ is a finite alphabet, an optimal solution to the integer program (3.2) can be found, surprisingly in $\mathcal{O}(|\mathcal{X}|\log(|\mathcal{X}|))$ time! This algorithm relies on the equivalence between prefix-free codes and binary trees, and presents a simple approach for constructing the code/binary trees. The scheme proceeds in the following steps:

- Sort the pmf to get $p_1 \geq p_2 \geq \ldots p_{m-1} \geq p_m$ for $m = |\mathcal{X}|$.
- Select the two symbols corresponding to p_{m-1} and p_m, and combine them into one symbol with probability $p_{m-1}+p_m$, and make this new symbol the parent node of the two original symbols.
- Repeat the process till only one symbol remains.

Example 96. Suppose $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and let X be an $\mathcal{X}$-valued random variable with pmf $(1/8, 1/8, 1/8, 1/8, 1/4, 1/4)$. In this case, the Huffman code exactly matches the entropy.

Theorem 97. *Huffman code is optimal for* (3.2). *That is, if C denotes the Huffman code, and C' is any other UD scheme, then $L_C(X) \leq L_{C'}(X)$ for all $\mathcal{X}$-valued random variables X.*

Despite its optimality, Huffman coding suffers from some drawbacks. The most important one is that in order to apply this scheme to large blocks of symbols (following Remark 95), the computational complexity grows exponentially. More specifically, even if we want to encode a single string of symbols in $\mathcal{X}^n$, we have to construct a code for all the $|\mathcal{X}|^n$ elements. We now present a scheme that can be constructed on-the-fly for streams of symbols in linear time.

Shannon-Fano-Elias (SFE) and Arithmetic Coding. The key idea behind the SFE coding scheme is to assign disjoint intervals to each symbol $x \in \mathcal{X} = \{1, 2, \dots, m\}$ using its cumulative distribution function (CDF). In particular, introduce the function $\overline{F}$ as

$$\overline{F}(x) = \sum_{x'<x} p(x') + \frac{1}{2}p(x) = F(x) - \frac{1}{2}p(x).$$

In words, $\overline{F}(x)$ denotes the mid-point of the interval corresponding to the jump in the CDF of X. We can then encode x using the first $l(x) = \lceil \log(1/p(x)) \rceil + 1$ bits of the binary representation of this value. Thus, the average codeword length of this scheme is within two bits of the entropy.

Now, suppose we want to apply this scheme to sequences of length n. Let us first order the elements of $\mathcal{X}^n$ based on the lexicographic order. A direct calculation of $\overline{F}(x^n)$ will have exponential complexity as before, since it requires summing up over all elements $y^n \leq x^n$. However, we can instead compute the CDF recursively in linear time using expression:

$$F_n(x^n) = F_{n-1}(x^{n-1}) + P_{X^{n-1}}(x^{n-1}) \sum_{x'<x} P_{X_n|X^{n-1}}(x').$$

This above approach has two advantages:

- It is an ‘anytime’ scheme: it does not require the blocklength to be fixed beforehand, as it can update the scheme on the fly.
- It is computationally efficient: the computation of $\overline{F}(x^n)$ is an $\mathcal{O}(n|\mathcal{X}|)$ operation.
- In the general case, where we do not know the probability distributions, it can be easily combined with any probability prediction schemes.

Finally, we end this section with the observation that all our discussion so far has focused on cases where we have perfect knowledge of the probability distribution. What happens under misspecification? That is, what if we assume that the distribution of X is q, when in reality, it is $p \neq q$?

Theorem 98. *Suppose we construct a code with lengths $l(x) = \lceil \log(1/q(x)) \rceil$. Then, we have*

$$H(P) + D_{kl}(P \parallel Q) \leq \mathbb{E}_P[l(X)] \leq H(P) + D_{kl}(P \parallel Q) + 1.$$

Thus, relative entropy has the operational meaning of the excess codeword length (or redundancy) incurred due to the lack of knowledge of the true distribution.

Proof. This result follows from a direct evaluation of $\mathbb{E}_P[l(X)]$. In particular, we have

$$\sum_{x\in\mathcal{X}} p(x) \log\left(\frac{1}{q(x)}\right) \leq \ \mathbb{E}_Q[l(X)] \ \leq \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{1}{q(x)}\right) + 1$$

On multiplying and dividing the term inside log with $p(x)$, we get the required statement:

$$\sum_{x\in\mathcal{X}} p(x) \left(\log\left(\frac{p(x)}{q(x)}\right) + \log\left(\frac{1}{p(x)}\right)\right) \leq \ \mathbb{E}_Q[l(X)] \ \leq \sum_{x\in\mathcal{X}} p(x) \left(\log\left(\frac{p(x)}{q(x)}\right) + \log\left(\frac{1}{p(x)}\right)\right) + 1.$$

□

3.4 Gambling

Consider the following scenario: there are m horses competing in a race, and the probability that horse $i \in \mathcal{X} = \{1, \ldots, m\}$ wins the race is equal to p_i. A bookmaker offers the odds of o_i-for-1 for each horse i, where $o_i \in [0, \infty)$. Let X denote the $\mathcal{X}$-valued random variable indicating the index of the winning horse. Suppose a gambler, with an initial wealth of $1, bets a fraction b_i on the i^{th} horse. Here $b_i \in [0,1]$, and $\sum_i b_i = 1$. We will use $\boldsymbol{b}$ and $\boldsymbol{o}$ to denote the betting and odds vectors, with $\boldsymbol{b}(i) = b_i$ and $\boldsymbol{o}(i) = o_i$ for $i \in \mathcal{X}$. Thus, the wealth of the gambler after betting on one horse race is equal to

$$W_1 = W_0 \times S(\boldsymbol{b}, \boldsymbol{o}, X) = 1 \times \boldsymbol{b}(X)\boldsymbol{o}(X).$$

We will refer to $S(\boldsymbol{b}, \boldsymbol{o}, X)$ as the *betting score*, and define the growth rate as

$$G(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{p}) = \mathbb{E}_{X \sim P}\left[\log S(\boldsymbol{b}, \boldsymbol{o}, X)\right] = \sum_{x \in \mathcal{X}} \boldsymbol{p}(x) \log\left(\boldsymbol{b}(x)\boldsymbol{o}(x)\right).$$

Our goal is to design betting strategies that maximize this objective.

Remark 99. The odds for a horse race can be specified in two ways: *a-for*-1 or *b-to-1*. Here are the steps involved in the former:

- Before the race, the bettor pays $1 to the bookmaker to bet on some event E (say horse i wins)
- If event E occurs, then the bettor receives $\$a$. Otherwise, the bettor gets nothing.

In the second description, all the exchanges happen after observing the event: prior to observing the event, the bettor puts $1 at stake. If the event E occurs, he receives $\$b$, otherwise he pays the bookmakers $1. It is easy to verify that the two descriptions are equivalent if $a = b + 1$.

Our first result establishes the log-optimal betting strategy for horse races.

Theorem 100. *For $X \sim P$ with a p.m.f. $\boldsymbol{p}$, the optimal growth rate, denoted by $A^* \equiv A(\boldsymbol{p}, \boldsymbol{o})$, is equal to*

$$G^* = \sum_{x \in \mathcal{X}} \boldsymbol{p}(x) \log(\boldsymbol{o}(x)) - H(\boldsymbol{p}). \tag{3.3}$$

Furthermore, this optimal value is achieved by a betting strategy $\boldsymbol{b} = \boldsymbol{p}$, and is known as the proportional betting *or* Kelly betting *strategy.*

Proof. We prove this in two steps: first we identify a candidate solution by differentiating the Lagrangian of this optimization problem, and then verifying its optimality using information theoretic inequalities.

TODO □

Corollary 101. *Suppose there are n i.i.d. horse races, with the same odds $\boldsymbol{o}$ in each race. Then, the optimal strategy is to bet according to $\boldsymbol{b}^* = \boldsymbol{p}$ in each round (i.e., a constant betting strategy), and the corresponding growth rate is nG^*, where G^* was defined in* (3.3)*.*

Example 102. Suppose there are two-horses with $\boldsymbol{p} = (p_1, p_2)$. If the bookmakers offer uniform odds of 2-for-1 on both the horses, the optimal betting strategy is the proportional betting strategy $\boldsymbol{b} = (p_1, p_2)$, and furthermore, the optimal growth rate is $G^* = \log(2) - h_2(\boldsymbol{p}) = 1 - h_2(\boldsymbol{p})$. It is easy to verify that G^* is equal to the capacity of a binary symmetric channel (BSC) with the flip probability p_1 (or equivalently $p_2 = 1 - p_1$).

Value of side-information. Suppose the bettor has some side-information Y taking values in some discrete set $\mathcal{Y}$, which he can use to bet on the outcome of the horse race X. Then, what is the financial value of this side-information? More specifically, how do the optimal betting strategy and the optimal growth rate change, given this side information.

Clearly, the performance of the bettor cannot be any worse than the case in which no side-information is available (since he can always choose to discard the side-information). Our next result gives a precise characterization of the utility of such side-information.

Proposition 103. *Suppose the gambler has access to some side information Y, and let P_{XY} denote the joint distribution of (X, Y). Then, the optimal growth rate with the side-information, denoted by G_s^*, is equal to*

$$G_s^* = G^* + I(X;Y).$$

Or in other words, the increase in the optimal growth rate is exactly equal to the mutual information between X and Y. Furthermore, the optimal betting strategy is the conditionally proportional betting strategy, that bets according to

$$\boldsymbol{b}^*(\cdot|Y=y) = \boldsymbol{p}(\cdot|Y=y).$$

Proof. For some conditional betting strategy $\boldsymbol{b}$, the growth rate is defined as

$$\begin{aligned} G &= \sum_{x,y} p(x,y)\log\left(o(x)b(x|y)\right) = \sum_{x,y} p(x,y)\log o(x) - \sum_{x,y} p(x,y)\log\left(\frac{1}{b(x|y)}\right) \\ &= \sum_{x,y} p(x,y)\log o(x) - \sum_{x,y} p(x,y)\log\left(\frac{p(x|y)}{b(x|y)}\right) + \sum_{x,y} p(x,y)\log\left(p(x|y)\right) \\ &= \sum_{x} p(x)\log o(x) - D_{\mathrm{kl}}(\boldsymbol{p}_{Y|X} \parallel \boldsymbol{b}_{Y|X}|\boldsymbol{p}_X) - H(Y|X). \end{aligned}$$

This implies that the optimal betting strategy is the conditional proportional betting strategy. Furthermore, the optimal growth rate with side-information is

$$\begin{aligned} G_s^* &= \sum_{x\in\mathcal{X}} p(x)\log o(x) - H(Y|X) == \sum_{x\in\mathcal{X}} p(x)\log o(x) - H(X) + H(X) - H(Y|X) \\ &= G^* + H(X) - H(X|Y) = G^* + I(X;Y). \end{aligned}$$

Thus, the mutual information between X and Y has the operational meaning as the financial value of side-information Y. □

The bookmaker's perspective. Suppose a bookmaker offers odds of o-for-1 for the occurrence of an event E (say a particular horse winning the race). Then, if the probability of E is equal to p, the expected returns for the bookmaker is

$$1 \times P(E^c) - (o-1) \times P(E) = (1-p) - (o-1)p = 1 - op.$$

Thus in order to make a profit, the bookmaker should offer odds satisfying $o \leq 1/p$. Or equivalently, we can consider $1/o$ as an upper bound on the bookmakers prediction about the probability of E.

Based on the above discussion, we will refer to betting odds as *fair*, *sub-fair*, and *super-fair*, depending on whether $\sum_{i=1}^m 1/o_i$ is equal to, larger than, or smaller than 1.

Proposition 104. *Suppose the betting odds offered for a horse race are fair; that is there exists an $\boldsymbol{r} = (r_1, \ldots, r_m) \in \Delta_m$, such that $r_i = 1/o_i$. Then, the growth rate for a strategy $\boldsymbol{b}$ is equal to*

$$G(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{p}) = D_{kl}(\boldsymbol{p} \parallel \boldsymbol{r}) - D_{kl}(\boldsymbol{p} \parallel \boldsymbol{b}), \quad \text{and} \quad G^* = D_{kl}(\boldsymbol{p} \parallel \boldsymbol{r}).$$

In words, a gambler can make money betting on horse races, only if the gamblers estimate ($\boldsymbol{b}$) of the probability of outcomes is closer to the truth ($\boldsymbol{p}$), than the bookmakers' estimate $\boldsymbol{r}$. The maximum growth rate achievable is equal to the divergence between the true probability and the bookmakers' estimate.

Proof. The proof follows directly from the definitions. In particular, we have

$$\begin{aligned} G \equiv G(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{p}) &= \sum_{x\in\mathcal{X}} p(x)\log(o(x)b(x)) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{b(x)}{r(x)}\right) \\ &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{b(x)p(x)}{r(x)p(x)}\right) = \sum_{x\in\mathcal{X}} p(x)\left(\log\left(\frac{p(x)}{r(x)}\right) - \log\left(\frac{p(x)}{b(x)}\right)\right) \\ &= D_{\mathrm{kl}}\left(\boldsymbol{p} \parallel \boldsymbol{r}\right) - D_{\mathrm{kl}}\left(\boldsymbol{p} \parallel \boldsymbol{b}\right). \end{aligned}$$

This implies that the optimal strategy is the one that minimizes the second term; that is $\boldsymbol{b} = \boldsymbol{p}$, which makes the second term zero. □

Remark 105. As mentioned earlier, the optimal betting strategy in this case is the proportional betting strategy $\boldsymbol{b}^* = \boldsymbol{p}^*$. Suppose a bettor makes a wrong prediction $\boldsymbol{q}$ about the probability of outcomes. Then, what is the amount of suboptimality due to his wrong prediction? The above result suggests that the cost of the ignorance of the true distribution is

$$\begin{aligned} G^* - G(\boldsymbol{q}) &= D_{\mathrm{kl}}(\boldsymbol{p} \parallel \boldsymbol{r}) - (D_{\mathrm{kl}}(\boldsymbol{p} \parallel \boldsymbol{r}) - D_{\mathrm{kl}}(\boldsymbol{p} \parallel \boldsymbol{q})) \\ &= D_{\mathrm{kl}}(\boldsymbol{p} \parallel \boldsymbol{q}). \end{aligned}$$

Thus, as in the case of compression, the relative entropy has an operational meaning as the price to pay for wrongly assuming the distribution to be $\boldsymbol{q}$, instead of the the true distribution($\boldsymbol{p}$).

Corollary 106. *If the odds on a horse race are uniform; that is $\boldsymbol{r} = (1/m, \ldots, 1/m)$, then we have the following 'conservation law':*

$$G^* + H(\boldsymbol{p}) = \log m.$$

Thus, a low entropy horse race (i.e., a more compressible or predictable source) can lead to a large rate of growth of the bettors wealth.

Proof. This result follows directly from Proposition 104, since

$$G^* = D_{\mathrm{kl}}(\boldsymbol{p} \parallel \boldsymbol{r}) = \log(m) + \sum_{x \in \mathcal{X}} p(x) \log p(x) = \log m - H(\boldsymbol{p}).$$

The result then follows by taking the entropy on to the left side of the equality. □

Betting with a 'cash' option. A subtle point in the preceding discussion is that the bettor is assumed to bet using all his money in each bet. In reality however, often the bettor might prefer to invest only a part of his wealth on the horse races. This can be modeled as betting with a 1-for-1 odds on a certain event. Such betting strategies can be represented as probability distributions over $\{0\} \cup \{1, \ldots, m\}$, where 0 denotes the cash option.

What happens to the optimal betting strategy when a cash option is allowed? Our next result addresses this issue.

Proposition 107. *Consider the problem of betting on horse races, where the bettor is allowed to retain a part of his initial wealth as cash. Then, the optimal betting strategy depends on the odds offered, and in particular:*

(a) If the betting odds are fair, or super-fair (i.e., $\sum_i 1/o_i$ is either $= 1$ or < 1), then the proportional betting strategy is still optimal. That is there is no incentive to set aside some of the money as cash.

(b) If the betting odds are sub-fair, then the optimal betting strategy is usually not known in a closed form, and can be found through numerical optimization.

Proof. We consider the super-fair and sub-fair cases separately.

(a) When the odds are either fair or super-fair, we will show that for any strategy that sets aside a part of the wealth as cash, i.e., bets with $\boldsymbol{b}(0) > 0$, there exists another strategy with $\boldsymbol{b}(0) = 0$, whose growth rate is at least as good. Hence, there is no loss in optimality by restricting to strategies that bet all the money; and we have proved that among such schemes, the optimal strategy is the proportional betting strategy.

In particular, suppose $\boldsymbol{b}$ denotes a strategy with $\boldsymbol{b}(0) > 0$. Then, consider another strategy $\boldsymbol{q}$, that is defined as follows:

$$\boldsymbol{q}(i) = \begin{cases} 0, & \text{if } i = 0 \\ \boldsymbol{b}(i) + \frac{c\boldsymbol{b}(0)}{\boldsymbol{o}(i)}, & \text{if } i \in \mathcal{X} = [m] \end{cases}, \quad \text{where} \quad c = \frac{1}{\sum_{i=1}^{m} 1/\boldsymbol{o}(i)}.$$

For the case of fair odds, the value of c is 1, while for super-fair odds, c is strictly greater than 1.

Now, for the outcome X, the betting score with the new strategy (with no cash option) is equal to

$$\begin{aligned}\boldsymbol{q}(X)\boldsymbol{o}(X) &= \left(\boldsymbol{b}(X) + \frac{c\boldsymbol{b}(0)}{\boldsymbol{o}(X)}\right)\boldsymbol{o}(X) = \boldsymbol{b}(X)\boldsymbol{o}(X) + c\boldsymbol{b}(0)\\ &\geq \boldsymbol{b}(X)\boldsymbol{o}(X) + \boldsymbol{b}(0).\end{aligned}$$

Thus the new strategy ($\boldsymbol{q}$) has the same gain as the strategy with cash option in the case of fair odds, and strictly better gain in the case of super-fair odds.

(b) In this case, we can write the following optimization problem:

$$\max_{\boldsymbol{b}} \sum_{i=1}^{m} p_i \log\left(b_0 + b_i o_i\right)$$

$$\text{subject to } \sum_{i=0}^{m} b_i = 1, \quad \text{and} \quad b_i \geq 0 \; \forall i \in \{0, \ldots, m\}.$$

For this optimization problem, we can write the Lagrangian as

$$L(\boldsymbol{b}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{m} p_i \log(b_0 + b_i o_i) + \lambda\left(\sum_{j=0}^{m} b_j\right) - \sum_{j=0}^{m} \mu_j b_j.$$

The first-order necessary conditions imply:

$$\begin{aligned}\frac{\partial L}{\partial b_0} &= \sum_{i=1}^{m} \frac{p_i}{b_0 + b_i o_i} + \lambda - \mu_0 = 0\\ \frac{\partial L}{\partial b_i} &= \frac{p_i o_i}{b_0 + b_i o_i} + \lambda - \mu_i = 0.\end{aligned}$$

The primal and dual constraints, along with the complementary slackness conditions, imply:

$$\sum_{i=0}^{m} b_i = 1, \quad b_i \geq 0 \; \forall i, \quad \lambda \geq 0, \quad \mu_i \geq 0, \quad \text{and} \quad \mu_i b_i = 0 \; \forall i.$$

By considering different cases, it can be proved that the optimal solution depends on the value of $A = \sum_{i=1}^{m} p_i o_i$. In particular, if $A \leq 1$, the the optimal strategy is to keep all the money as cash: that is $\boldsymbol{b}(0) = 1$.

In the case where $A = \sum_{i=1}^{m} p_i o_i > 1$, the optimal strategy cannot be explicitly stated, but can be algorithmically constructed, and we refer the reader to Kelly [1956] for details.

□

Repeated betting on horse races. If there is a sequence of n horse races, with $X^n \in \mathcal{X}^n$ indicating the winners of these races, we can similarly define the wealth after n rounds as

$$W_n = \boldsymbol{b}^n(X^n)\boldsymbol{o}^n(X^n) = \prod_{i=1}^{n} \boldsymbol{b}(X_i|X^{i-1})\boldsymbol{o}(X_i|X^{i-1}).$$

The above definition allows both, the betting strategy and the odds offered, to be predictable: that is, for the i^{th} race, the betting method $\boldsymbol{b}(\cdot|X^{i-1})$ and the odd $\boldsymbol{o}(\cdot|X^{i-1})$ can depend on the outcomes of the prior races. It also allows the horse races to be dependent on each other. Theorem 100 implies that the growth optimal betting strategy in this setting is to $\boldsymbol{b}^n = \boldsymbol{p}^n$.

Remark 108. The above discussion makes the connection between gambling and compression clear: both these problems can be reduced to that of assigning probabilities to sequences of observations. In the case of compression, by assigning probabilities, we can use methods such as arithmetic coding, to convert them into a prefix-free code. While in gambling, by assigning probabilities to all sequences, we can increase the rate at which the bettor's wealth grows.

3.5 Portfolio Optimization

Consider a stock-market with m stocks and let $\boldsymbol{X} = (X_1, \ldots, X_m)$ denote the price relative of these stocks over one trading period (say one day). That is, X_i denotes the price ratio of the price of stock $i \in [m]$ at the end of the day to its price at the beginning of the day. We assume that $\boldsymbol{X}$ is drawn according to some distribution P, defined on the set $[0, \infty)^m$.

A stock-portfolio is represented by a vector $\boldsymbol{b} \in \Delta_{m-1}$, that denotes the distribution of the investment over the m stocks. Starting with a wealth of \$1, the wealth after one day for a portfolio $\boldsymbol{b}$ is defined as

$$W_1 = W_0 \times \langle \boldsymbol{X}, \boldsymbol{b} \rangle = W_0 \times \left(\sum_{i=1}^{m} X_i b_i \right).$$

Often, one of the components of a portfolio is a risk-free asset, such as treasury bonds or cash. Such assets correspond to a price relative of 1.

Remark 109. The horse races from the previous section are a special case of the general stock-market introduced above. In particular, it corresponds to the case where $\boldsymbol{X}$ is restricted to take values in the discrete set $\{o_1 e_1, \ldots, o_m e_m\}$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ denotes the element in $\{0,1\}^m$, with 1 at the i^{th} coordinate. Thus, the stock market is a significant generalization of horse races, allowing $\boldsymbol{X}$ to take any values in $[0, \infty)^m$. However, as we will see in the next chapter, in some sense, horse races are the examples of the *hardest* stock markets.

Remark 110. Consider the setting where an *expert* recommends us to use a particular portfolio $\boldsymbol{b}$. Since the components of X are not assumed to be independent, we can define a new price relative $Y = (X, \langle \boldsymbol{b}, X \rangle)$. So, do we gain in terms of the growth rate by expanding the portfolio of stocks and betting on this larger portfolio? The answer is no; that is,

$$\sup_{\widetilde{\boldsymbol{b}} \in \Delta_m} \mathbb{E}_Y[\log \langle \widetilde{\boldsymbol{b}}, Y \rangle] = \sup_{\boldsymbol{b} \in \Delta_{m-1}} \mathbb{E}_X \left[\log \langle \boldsymbol{b}, X \rangle \right].$$

The growth rate of a stock portfolio $\boldsymbol{b}$, and the optimal growth rate are defined as

$$G(\boldsymbol{b}, P) = \mathbb{E}_P \left[\log \langle \boldsymbol{X}, \boldsymbol{b} \rangle \right], \quad G^*(P) = \sup_{\boldsymbol{b} \in \Delta_{m-1}} G(\boldsymbol{b}, P).$$

Unlike the horse racing problem, the optimal portfolio $\boldsymbol{b}^*$ does not admit a closed-form expression in general. Nevertheless, we can establish necessary and sufficient conditions for a $\boldsymbol{b}$ to be optimal.

Theorem 111. *The optimal portfolio $\boldsymbol{b}^*$ for a stock market $\boldsymbol{X} \sim P$ satisfies*

$$\mathbb{E}_P \left[\frac{X_i}{\langle \boldsymbol{b}^*, \boldsymbol{X} \rangle} \right] = \begin{cases} = 1, & \text{if } b_i^* > 0 \\ \leq 1 & \text{if } b_i^* = 0. \end{cases}$$

Proof. To begin proving this, we first observe the following two facts:

- The mapping $\boldsymbol{b} \mapsto G(\boldsymbol{b}, P) := \mathbb{E}_P \left[\log \langle \boldsymbol{b}, \boldsymbol{X} \rangle \right]$ is concave. This is because, $\boldsymbol{b} \mapsto \langle \boldsymbol{X}, \boldsymbol{b} \rangle$ is linear, $x \mapsto \log(x)$ is concave, and that $\mathbb{E}_P[\cdot]$ is a linear operator.

- The mapping $P \mapsto G^*(P)$ is convex. To see why this is true, consider distributions P_1, P_2, and their convex combinations P_λ. Then, we have

$$\begin{aligned} G^*(P_\lambda) &= G\left(\boldsymbol{b}_\lambda^*, P_\lambda\right) = \lambda G\left(\boldsymbol{b}_\lambda^*, P_1\right) + (1-\lambda) G\left(\boldsymbol{b}_\lambda^*, P_2\right) \\ &\leq \lambda G^*(P_1) + (1-\lambda) G^*(P_2). \end{aligned}$$

- The set of log-optimal portfolios for a fixed P is convex. To see this, consider two log-optimal portfolios $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. Then, due to the concavity of the mapping from $\boldsymbol{b} \mapsto G(\boldsymbol{b}, P)$, we have

$$G(\boldsymbol{b}_\lambda, P) \geq \lambda G(\boldsymbol{b}_1, P) + (1-\lambda) G(\boldsymbol{b}_2, P) = G^*(P).$$

  Hence, $\boldsymbol{b}_\lambda$ is also log-optimal.

Thus, the problem of finding a log-optimal portfolio reduces to maximizing a concave function over a convex set, and the set of optimal solutions is also convex. Let $\boldsymbol{b}^*$ denote any log-optimal portfolio. Then, by the first-order necessary condition for optimality, the directional derivative along a direction $\boldsymbol{b}$ must be nonnegative. That is

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left(\mathbb{E}_P \left[\log \langle (1-\lambda)\boldsymbol{b}^* + \lambda \boldsymbol{b}, X \rangle \right] - \mathbb{E}_P \left[\log \langle \boldsymbol{b}^*, X \rangle \right] \right) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left(\mathbb{E}_P \left[\log \left(\frac{\langle (1-\lambda)\boldsymbol{b}^* + \lambda \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} \right) \right] \right) \leq 0.$$

On simplification, the above result implies

$$\begin{aligned}
\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left(\mathbb{E}_P \left[\log \left(\frac{\langle (1-\lambda)\boldsymbol{b}^* + \lambda \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} \right) \right] \right) &= \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left(\mathbb{E}_P \left[\log \left(1 - \lambda + \lambda \frac{\langle \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} \right) \right] \right) \\
&\overset{DCT}{=} \mathbb{E}_P \left[\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \log \left(1 + \lambda \left(\frac{\langle \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} - 1 \right) \right) \right] \\
&= \mathbb{E}_P \left[\left(\frac{\langle \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} - 1 \right) \right]
\end{aligned}$$

Thus, the above discussion implies that

$$\mathbb{E}_P \left[\frac{\langle \boldsymbol{b}, X \rangle}{\langle \boldsymbol{b}^*, X \rangle} \right] \leq 1.$$

By setting $b = e_i$, we get $\mathbb{E}_P \left[X_i / \langle \boldsymbol{b}^*, X \rangle \right] \leq 1$. Furthermore, if $b_i^* > 0$, the inequality holds with equality. □

Consequences of the characterization of log-optimal portfolio. Using the previous result, we can derive several interesting properties of the log-optimal portfolio. In particular, we will use the above result to show the following:

- The expected ratio of the one-step increments of any portfolio to a log-optimal portfolio is at most one.
- Playing the log-optimal portfolio repeatedly results in a wealth that is at least as good as any other predictable scheme in the first order term in the exponent.
- The drop in optimal growth rate due to using a wrong distribution for X (say Q instead of P) is upper bounded by the relative entropy $D_{\mathrm{kl}}(P \parallel Q)$.
- The value of side-information Y in portfolio optimization is upper bounded by the mutual information $I(X;Y)$.

We begin with a result that states that the ratio of the one-step growth of any portfolio $\boldsymbol{b}$ over that of the optimal portfolio is smaller than 1 in expectation.

Corollary 112. *Suppose $S^* = \langle \boldsymbol{b}^*, X \rangle$, and $S = \langle \boldsymbol{b}, X \rangle$ for any feasible portfolio $\boldsymbol{b}$. Then, we have*

$$\mathbb{E}_P \left[\log \left(\frac{S}{S^*} \right) \right] \leq 0 \quad \Leftrightarrow \quad \mathbb{E}_P \left[\frac{S}{S^*} \right] \leq 1.$$

As a result, we have

$$\mathbb{P} \left(\frac{S}{S^*} \geq \delta \right) \leq \frac{1}{\delta},$$

by an application of Markov's inequality.

Proof. Suppose $\boldsymbol{b}^*$ is a log-optimal portfolio. Then, by Theorem 111, we have

$$\mathbb{E}_P \left[\frac{X_i}{S^*} \right] \leq 1,$$

which implies that

$$\sum_{i=1}^{m} b_i \mathbb{E}_P\left[\frac{X_i}{S^*}\right] = \sum_{i=1}^{m} \mathbb{E}_P\left[\frac{b_i X_i}{S^*}\right] = \mathbb{E}_P\left[\frac{S}{S^*}\right] \leq \sum_{i=1}^{m} b_i = 1.$$

This proves one direction: if $\boldsymbol{b}^*$ is the log-optimal portfolio, then the ratio one-step increments of any other strategy w.r.t. $\boldsymbol{b}^*$ is upper bounded by 1 in expectation.

To show the other direction, that if $\mathbb{E}_P[S/S^*] \leq 1$ for some strategy $\boldsymbol{b}^*$, then it is log-optimal, we simply note that

$$\mathbb{E}_P\left[\log\left(\frac{S}{S^*}\right)\right] \leq \log\left(\mathbb{E}_P\left[\frac{S}{S^*}\right]\right) \leq \log(1) = 0.$$

□

Remark 113. Consider a problem with two stocks: $(X_1, X_2) = (1, 1/(1-\epsilon))$ or $(1, 0)$, with probability $1-\epsilon$ and ϵ respectively. How does the optimal portfolio look like? Suppose $\boldsymbol{b} = (b, \bar{b})$. Then, applying Theorem 111 to this problem, we have

$$\mathbb{E}\left[\frac{X_2}{bX_1 + \bar{b}X_2}\right] = (1-\epsilon)\frac{1/(1-\epsilon)}{b + \bar{b}/(1-\epsilon)} = \frac{1-\epsilon}{1-\epsilon b} \leq 1.$$

This implies that $b \geq 1$, which means that the optimal portfolio is $\boldsymbol{b}^* = (1, 0)$, and it places no weight on the second stock.

Now, consider another investor, who invests according to the opposite portfolio $\boldsymbol{b} = (0, 1)$, and invests all his money in the second stock. In one round of investment, the second investor outperforms the log-optimal investor with probability at least $1 - \epsilon$. Thus, the inequality in Corollary 112 is tight in this sense.

As in the case of horse racing, we can consider the problem of repeated investments in stock markets. In such problems, we can adapt the betting strategy based on the previous outcomes. However, for i.i.d. stock markets, the optimal strategy is a constant investment strategy (i.e., the one-step log-optimal strategy).

Proposition 114. *Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote i.i.d. realizations of P. If W_n^* is the wealth of the log-optimal strategy, then we have*

$$\mathbb{E}_{P^n}[\log(W_n^*)] = nG^*(P) \geq \mathbb{E}_{P^n}[\log W_n],$$

for any W_n constructed using predictable investment strategies.

Proof. The proof of the above statement follows from the definitions.

$$\max_{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n} \mathbb{E}\left[\log W_n\right] = \max_{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n} \sum_{i=1}^{n} \log\langle \boldsymbol{b}_i, \boldsymbol{X}_i\rangle = n \max_{\boldsymbol{b}} \mathbb{E}\left[\log\langle \boldsymbol{b}, \boldsymbol{X}\rangle\right] = nG^*(P).$$

□

Using this, we can show that playing the log-optimal portfolio repeatedly is optimal up to the first exponent.

Proposition 115. *Let W_n^* denote the wealth from playing the log-optimal strategy repeatedly. Then, for any other predictable strategy, we have*

$$\limsup_{n\to\infty} \log\left(\frac{W_n}{W_n^*}\right) \leq 0 \quad \textit{almost surely.}$$

Proof. This result is a simple consequence of Corollary 112 and the first Borel-Cantelli Lemma. In particular, due to Corollary 112, we have

$$\mathbb{P}\left(\frac{W_n}{W_n^*} \geq a_n\right) \leq \frac{\mathbb{E}\left[W_n/W_n^*\right]}{a_n} \leq \frac{1}{a_n}.$$

Now, if we select $a_n = n^2$, it implies that $\sum_n 1/a_n < \infty$. Hence, by an application of Borel-Cantelli lemma, we have

$$\mathbb{P}\left(\frac{1}{n}\log(W_n/W_n^*) > \frac{1}{n}\log(a_n) \text{ infinitely often}\right) = 0.$$

Since $\log(a_n)/n \to 0$, this implies that

$$\limsup_{n\to\infty} \frac{1}{n}\log\left(\frac{W_n}{W_n^*}\right) \leq \lim_{n\to\infty} \frac{2\log n}{n} = 0$$

almost surely. This completes the proof. □

While the constantly rebalanced portfolio b^* is optimal for the case of i.i.d. price relatives, in general, we the optimal strategy is the conditionally log-optimal portfolios. That is $\{X_t : t \geq 1\}$ is an arbitrary stochastic process, then the optimal strategy is to play the portfolio $\boldsymbol{b}_t^* \equiv \boldsymbol{b}_t^*(\cdot|X^{t-1})$, based on the conditional distribution of X_t given X^{t-1}.

We now show an analog of Corollary 112 for this general scheme.

Proposition 116. *Let $\{X_t : t \geq 1\}$ denote a sequence of price relatives drawn according to an arbitrary stochastic process. Let $\{W_n^* : n \geq 1\}$ and $\{W_n : n \geq 1\}$ denote the wealth processes of the conditionally log-optimal investor, and an arbitrary predictable investor. Then, we have the following:*

(a) The ratio process $\{M_n = W_n/W_n^* : n \geq 1\}$ is a nonnegative supermartingale adapted to the natural filtration $\{\mathcal{F}_n = \sigma(X^n) : n \geq 1\}$, with an initial value 1. (i.e., it is a test martingale*)*

(b) There exists a $\mathcal{F}_\infty$-measurable random variable M, such that $M_n \xrightarrow{a.s.} M$, and $\mathbb{E}[M] \leq 1$.

(c) For any $\delta > 0$, we have

$$\mathbb{P}\left(\exists n \geq 1 : M_n \geq \frac{1}{\delta}\right) \leq \delta.$$

Proof. The main step in this proof is to show the supermartingale property.

(a) In particular, we observe that

$$\begin{aligned}\mathbb{E}\left[\frac{W_{n+1}}{W_{n+1}^*} \mid \mathcal{F}_n\right] &= \mathbb{E}\left[\frac{W_n\langle \boldsymbol{b}_{n+1}, X_{n+1}\rangle}{W_n^*\langle \boldsymbol{b}_{n+1}^*, X_{n+1}\rangle} \mid \mathcal{F}_n\right] \\ &= \frac{W_n}{W_n^*}\sum_{i=1}^m \boldsymbol{b}_{n+1}(i)\mathbb{E}\left[\frac{X_{n+1}(i)}{\langle \boldsymbol{b}_{n+1}^*, X_{n+1}\rangle} \mid \mathcal{F}_n\right] \\ &\leq M_n \sum_{i=1}^m \boldsymbol{b}_{n+1}(i) = M_n.\end{aligned}$$

The nonnegativity follows from the definitions of W_n and W_n^*. Thus, we have proved that $\{M_n : n \geq 0\}$ is a nonnegative supermartingale with an initial value of 1.

(b) These statements are a direct application of the (super-)Martingale convergence theorem.

(c) This is an application of Ville's inequality.

□

Finally, we consider the issues of misspecified probability distribution, and the value of side information.

Proposition 117. *Suppose $X \sim P$, and let $\boldsymbol{b}_P$ denote the optimal portfolio for this distribution over price relatives. Let Q denote another distribution over price relatives, and $\boldsymbol{b}_Q$ its corresponding optimal portfolio. Then, if $X \sim P$, but an investor wrongly assumes that the distribution is Q, the reduction in growth rate satisfies*

$$G^*(P) - G(\boldsymbol{b}_Q, P) \leq D_{kl}(P \parallel Q).$$

Proof. The proof of this result follows the definition along with the characterization of the optimal portfolio in Theorem 111. For simplicity, we will assume that P and Q have densities p and q with respect to the Lebesgue measure on $\mathbb{R}^m$. Then, we have the following:

$$\begin{aligned} G^*(P) - G(\boldsymbol{b}_Q, P) &= \mathbb{E}_P\left[\log\langle \boldsymbol{b}_P, X\rangle\right] - \mathbb{E}_P\left[\log\langle \boldsymbol{b}_Q, X\rangle\right] = \mathbb{E}_P\left[\log\left(\frac{\langle \boldsymbol{b}_P, X\rangle}{\langle \boldsymbol{b}_Q, X\rangle}\right)\right] \\ &= \mathbb{E}_P\left[\log\left(\frac{\langle \boldsymbol{b}_P, X\rangle q(X)}{\langle \boldsymbol{b}_Q, X\rangle p(X)}\right) + \log\left(\frac{p(X)}{q(X)}\right)\right]. \end{aligned}$$

Now, the expectation of the second term is clearly the relative entropy between P and Q, and to complete the proof, we need to show that the expectation of the first term is nonnegative. We do this by appealing to Theorem 111, and concavity of $\log(\cdot)$. In particular, we have

$$\begin{aligned} G^*(P) - G(P, \boldsymbol{b}_Q) &= D_{\mathrm{kl}}(P \parallel Q) + \mathbb{E}_P\left[\log\left(\frac{\langle \boldsymbol{b}_P, X\rangle q(X)}{\langle \boldsymbol{b}_Q, X\rangle p(X)}\right)\right] \\ &\leq D_{\mathrm{kl}}(P \parallel Q) + \log \mathbb{E}_P\left[\frac{\langle \boldsymbol{b}_P, X\rangle q(X)}{\langle \boldsymbol{b}_Q, X\rangle p(X)}\right] \\ &= D_{\mathrm{kl}}(P \parallel Q) + \log \mathbb{E}_Q\left[\frac{\langle \boldsymbol{b}_P, X\rangle}{\langle \boldsymbol{b}_Q, X\rangle}\right] \\ &= D_{\mathrm{kl}}(P \parallel Q) + \log\left(\sum_{i=1}^m \boldsymbol{b}_P(i)\mathbb{E}_Q\left[\frac{X_i}{\langle \boldsymbol{b}_Q, X\rangle}\right]\right) \\ &\leq D_{\mathrm{kl}}(P \parallel Q) + \log\left(\sum_{i=1}^m \boldsymbol{b}_P(i)\right) = D_{\mathrm{kl}}(P \parallel Q). \end{aligned}$$

The first inequality above is due to an application of Jensen's inequality, and the second inequality uses the characterization of the optimal portfolio $\boldsymbol{b}_Q$. □

A simple application of the previous result is that the value of side-information is upper bounded by the mutual information, with equality if the stock market is actually a horse race.

Corollary 118. *Suppose Y denotes some side-information that is available to the investor. Then, the improvement in the growth rate by incorporating the side-information is no larger than the mutual information $I(X;Y)$. That is,*

$$G_s^*(P) - G^*(P) \leq I(X;Y).$$

The equality holds when the stock market is actually a horse race.

Proof. To prove this result, we again look at the definition of the drop in growth rates.

$$\begin{aligned} G_s^*(P) - G^*(P) &= \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[\log\langle \boldsymbol{b}_Y^*, X\rangle | Y]\right] - \mathbb{E}_X\left[\log\langle \boldsymbol{b}^*, X\rangle\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{X|Y}\left[\log\left(\frac{\langle \boldsymbol{b}_Y^*, X\rangle}{\langle \boldsymbol{b}^*, X\rangle}\right) \mid Y\right]\right] \\ &\leq \mathbb{E}_Y\left[\mathbb{E}_{X|Y}\left[\log\left(\frac{p_{X|Y}(X)}{p_X(X)}\right)\right]\right] \\ &= D_{\mathrm{kl}}\left(P_{X|Y} \parallel P_X | P_Y\right) = I(X;Y), \end{aligned}$$

where the inequality is due to Proposition 117. □

Chapter 4

Universal Compression and Gambling

We consider two versions of the universal compression problem:

- Setting I: probabilistic setting, where the observations are drawn from an unknown distribution P from a known class $\mathcal{P}$. The goal is to develop a compression scheme that minimizes the worst case (over all $P \in \mathcal{P}$) redundancy.
- Setting II: individual sequence setting, where our goal is to compress sequences of observations with no probabilistic assumptions on the source. The objective is to compress as well as the best distribution from some comparator class $\mathcal{P}$.

4.1 Universal Compression I: Probabilistic Setting

Let $\mathcal{P}$ denote a class of distributions on an alphabet $\mathcal{X}$. Our goal is to design a compressor Q that minimizes the worst case redundancy. In particular, we recall that the redundancy of Q with respect to the distribution P is defined as

$$R(Q, P) = \mathbb{E}_P \left[\log\left(\frac{1}{q(X)}\right) - \log\left(\frac{1}{p(X)}\right) \right] = D_{\mathrm{kl}}(P \parallel Q).$$

Now, the worst case redundancy for Q, and the minimax optimal value of redundancy for the class $\mathcal{P}$ are defined as follows:

$$R(Q, \mathcal{P}) = \max_{P \in \mathcal{P}} D_{\mathrm{kl}}(P \parallel Q), \quad \text{and} \quad R^*(\mathcal{P}) = \min_Q \max_{P \in \mathcal{P}} D_{\mathrm{kl}}(P \parallel Q) = \min_Q R(Q, \mathcal{P}). \tag{4.1}$$

Remark 119. Note that, for most non-trivial distribution classes $\mathcal{P}$, in order to achieve a small redundancy, we need to allow Q to be an arbitrary distribution, and more specifically, it should not be restricted to lie in $\mathcal{P}$. For example, consider the product space $\mathcal{X}^n$, and let $\mathcal{P}$ denote the class of all i.i.d. distributions on $\mathcal{X}^n$. Then, if Q is restricted to lie in $\mathcal{P}$, then it is easy to verify that $R^*(\mathcal{P}) = \Omega(n)$. That is, it is not possible to achieve sublinear redundancy without allowing Q to lie outside $\mathcal{P}$.

Our goal is to design schemes that match the optimum redundancy $R^*(\mathcal{P})$ in a computationally efficient manner. First, we present a simple typical-set based compression scheme, that achieves sup-linear redundancy for the class of i.i.d. distributions on finite alphabet $\mathcal{X}$.

4.1.1 Warmup: Typicality based compression

Let $\mathcal{X} = \{x_1, \ldots, x_m\}$ denote a finite alphabet. For any $n \geq 1$, and $x^n \in \mathcal{X}^n$, define the following terms:

- Let $\widehat{P}_{x^n}$ denote the "type" or the "empirical distribution" associated with x^n. Let $\mathcal{P}_n$ denote the set of all distinct types possible using elements of $\mathcal{X}^n$. It is easy to verify that $|\mathcal{P}_n| = \binom{n+m-1}{m-1} \leq (n+1)^m$. Let K_n denote the cardinality of $\mathcal{P}_n$, and furthermore, enumerate the possible types $\widehat{P}_1, \ldots, \widehat{P}_{K_n}$.

- For every $\widehat{P} \in \mathcal{P}_n$, let $T(\widehat{P})$ denote the "type class" of $\widehat{P}$. That is, $T(\widehat{P})$ represents all the sequences $x^n \in \mathcal{X}^n$, such that $\widehat{P}_{x^n} = \widehat{P}$. Recall that we know the following

$$\frac{2^{nH(\widehat{P})}}{K_n} \;\leq\; |T(\widehat{P})| \;\leq\; 2^{nH(\widehat{P})}.$$

  For every type class, enumerate all the elements in some order.

We are now ready to describe the coding scheme for all i.i.d. sources on the finite alphabet $\mathcal{X}$.

Definition 120 (Type-based universal code)**.** Given a sequence $x^n \in \mathcal{X}^n$, proceed as follows:

- Identify the type $\widehat{P} \equiv \widehat{P}_{x^n}$ of the sequence.
- Encode the index of the type of $\widehat{P}_{x^n}$ among the K_n possible types, using $\lceil \log K_n \rceil$ bits. Denote this by $C_1(x^n)$
- Next, identify the position of the sequence x^n in the type class $T(\widehat{P}_{x^n})$.
- Encode this index using $\lceil \log |T(\widehat{P}_{x^n})| \rceil$ bits. Denote this with $C_2(x^n)$.
- Concatenate the two to get the codeword for x^n. That is $C(x^n) = C_1(x^n)C_2(x^n)$.

The main result of this section is that the redundancy of this coding scheme is sublinear in the blocklength (i.e., n), for any i.i.d. distribution P^n on $\mathcal{X}^n$.

Proposition 121. *Let $X^n \sim P^n$ denote an i.i.d. sequence on $\mathcal{X}^n$. Let $C(X^n)$ denote the codeword assigned to X^n by the scheme described in Definition 120. Then, we have*

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{P^n}[\ell_C(X^n)] = 0,$$

for all $P \in \mathcal{P}(\mathcal{X})$.

Proof. The proof of this result uses standard typicality arguments. In particular, fix an arbitrary $\epsilon > 0$, and define the typical set $A_\epsilon = \{x^n \in \mathcal{X}^n : |-\log(p^n(x^n))/n - H(P)| \leq \epsilon\}$. Then, for n large enough, recall the following facts:

- $P^n(A_\epsilon) \geq 1-\epsilon$.
- $2^{-n(H(P)+\epsilon)} \leq P^n(x^n) \leq 2^{-n(H(P)-\epsilon)}$.
- $2^{n(H(P)-\epsilon)} \leq |A_\epsilon| \leq 2^{n(H(P)+\epsilon)}$.

Now, consider any $x^n \in A_\epsilon \cap T(\widehat{P})$ with type $\widehat{P}$. We know that the probability assigned to it by P^n satisfies

$$2^{-n(H(P)+\epsilon)} \leq P^n(x^n) = 2^{-n\left(H(\widehat{P})+D_{\mathrm{kl}}(\widehat{P}\|P)\right)} \leq 2^{-n(H(P)-\epsilon)}.$$

This, in turn, implies

$$H(\widehat{P}) + D_{\mathrm{kl}}(\widehat{P} \parallel P) \leq H(P) + \epsilon, \quad \Rightarrow \quad H(\widehat{P}) \leq H(P) + \epsilon.$$

Hence, we have

$$A_\epsilon \subset B_\epsilon := \bigcup_{\widehat{P}:H(\widehat{P})\leq H(P)+\epsilon} T(\widehat{P}).$$

Clearly, $P^n(B_\epsilon) \geq P^n(A_\epsilon) \geq 1-\epsilon$. Now, note the following:

- The number of bits assigned to elements of B_ϵ is no larger than $n(H(P)+\epsilon) + m\log(n+1) + 2$.

- The number of bits assigned to elements of $\mathcal{X}^n \setminus B_\epsilon$ is trivially upper bounded by $n \log m + m \log(n + 1) + 2$.

These two facts imply that the expected codeword length is upper bounded by

$$\mathbb{E}_P[\ell_C(X^n)] \leq (1-\epsilon)\left(nH(P) + n\epsilon + 2 + m\log(n+1)\right) + \epsilon\left(n\log m + m\log(n+1) + 2\right).$$

Or, in other words, we have

$$\frac{1}{n}\mathbb{E}_P[\ell_C(X^n)] \leq (1-\epsilon)H(P) + \epsilon\left(1 + \log m\right) + \frac{m\log(n+1)+2}{n}.$$

Taking the limit to n, we get the required statement

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}_P[\ell_C(X^n)] \leq H(P) + \epsilon\left(1 + \log m - H(P)\right) = H(P) + \mathcal{O}(\epsilon).$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

The above construction shows that there exist coding schemes that asymptotically achieve the optimal compression rate for any i.i.d. distribution on a per symbol basis. We now investigate what is the correct convergence rate.

4.1.2 Redundancy-Capacity

We now return to the analysis of the minimax redundancy, defined in (4.1). For simplicity, we focus on the case of a finite class of distributions, $\mathcal{P} = \{P_1, \ldots, P_N\}$ on some discrete alphabet $\mathcal{X}$. Note that this $\mathcal{X}$ itself could be some product space (say $\mathcal{Y}^n$), and the $P_i's$ might represent the distributions over these products spaces.

To state the main result of this section, we need to introduce some notation. Let U be any $[N]$-valued random variable with distribution π_U. Define the channel $P_{X|U}$ with $P_{X|U=j} = P_j$. Then, the capacity of this channel is defined as

$$C = \sup_{\pi_U} I(U;X). \tag{4.2}$$

Our main result of this section connects this term to the minimax redundancy.

Theorem 122. *The minimax redundancy for the class $\mathcal{P}$ is equal to the capacity defined in (4.2). That is,*

$$R^*(\mathcal{P}) = \min_Q \max_{P_i \in \mathcal{P}} D_{kl}(P_i \parallel Q) = C = \sup_{\pi_U} I(U;X).$$

Furthermore, the optimal coding distribution is $Q = Q_{\pi^*} = \sum_{i=1}^n \pi^*(i)P_i$, where π^* is the capacity achieving input distribution.

Proof. The proof of this result proceeds in two steps:

- The first step is to use the minimax theorem to observe that

$$\min_Q \max_{P_i} D_{\mathrm{kl}}(P_i \parallel Q) = \max_\pi \min_Q \sum_{i=1}^N \pi_i D_{\mathrm{kl}}(P_i \parallel Q).$$

- The second step is to show that $\min_Q \sum_{i=1}^N \pi_i D_{\mathrm{kl}}(P_i \parallel Q)$ is actually equal to $I(U;X)$ with $X \sim \pi$.

To see the first result, we note that

$$\min_Q \max_{P_i \in \mathcal{P}} D_{\mathrm{kl}}(P_i \parallel Q) = \min_Q \max_\pi \sum_{i=1}^N \pi_i D_{\mathrm{kl}}(P_i \parallel Q) = \min_Q \max_\pi \mathcal{C}(Q,\pi).$$

The objective function $\mathcal{C}$ is convex in the first argument, and linear (hence concave) in the second argument. Furthermore, both the action spaces are compact. Thus, by an application of the minimax theorem, we have

$$R^*(\mathcal{P}) = \max_{\pi} \ \min_{Q} \ \mathcal{C}(Q, \pi) = \max_{\pi} \ \min_{Q} \ \sum_{i=1}^{N} \pi_i D_{\mathrm{kl}}(P_i \parallel Q).$$

To complete the proof, we will show the variational definition of the mutual information between U and X.

$$\begin{aligned} I(U;X) &= D_{\mathrm{kl}}(P_{UX} \parallel \pi P_X) = D_{\mathrm{kl}}\left(P_{X|U} \parallel P_X|\pi\right) \\ &= \sum_{i=1}^{N} \pi_i D_{\mathrm{kl}}(P_i \parallel P_\pi) = \sum_{i=1}^{N} \pi_i \sum_{x \in \mathcal{X}} p_i(x) \log\left(\frac{p_i(x)}{p_\pi(x)}\right). \end{aligned}$$

Now consider any distribution Q over $\mathcal{X}$ with p.m.f. q. Then, multiplying and dividing inside the logarithm with $q(x)$, we get

$$\begin{aligned} I(U;X) &= \sum_{i=1}^{N} \pi_i \sum_{x \in \mathcal{X}} p_i(x) \left(\log\left(\frac{p_i(x)}{q(x)}\right) - \log\left(\frac{p_\pi(x)}{q(x)}\right)\right) \\ &= \sum_{i=1}^{n} \pi_i D_{\mathrm{kl}}(P_i \parallel Q) - \sum_{x \in \mathcal{X}} \sum_{i=1}^{n} \pi_i p_i(x) \log\left(\frac{p_\pi(x)}{q(x)}\right) \\ &= \sum_{i=1}^{n} \pi_i D_{\mathrm{kl}}(P_i \parallel Q) - D_{\mathrm{kl}}(P_\pi \parallel Q) \end{aligned}$$

Thus, we have proved that

$$I(U;X) = \min_{Q} \sum_{i=1}^{n} \pi_i D_{\mathrm{kl}}(P_i \parallel Q),$$

with equality achieved at $Q = P_\pi$, where $U \sim \pi$. THe final step is to observe that

$$C = \max_{\pi} I(U;X) = \max_{\pi} \min_{Q} \sum_{i=1}^{N} \pi_i D_{\mathrm{kl}}(P_i \parallel Q) = R^*(\mathcal{P}).$$

This completes the proof. □

Some remarks on this result:

- Our goal was to find a single distribution Q that minimizes the worst case redundancy against an adversarially chosen $P_i \in \mathcal{P}$. This result tells us that the optimal Q is a convex combination (i.e., a mixture) of all the distributions in $\mathcal{P}$. Furthermore, surprisingly it also says that the optimal mixing distribution is the same as the c.a.i.d. of the channel $P_{X|U}$ introduced earlier.
- The result is also true more generally for arbitrary finite dimensional parametric families of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ for some finite $d \geq 1$.

Connection to exponentially weighted predictors. Consider the n-fold product of a finite alphabet, denoted by $\mathcal{X}^n$, and let $\{p_\theta : \theta \in \Theta\}$ denote a class of distributions over $\mathcal{X}^n$. For any distribution π over the parameter set Θ, let q^n the corresponding mixture distribution over $\mathcal{X}^n$. Then, we have

$$q^n(x^n) = \int_\Theta \pi(\theta) p^n(x^n) d\theta = \prod_{i=1}^{n} q_i(x_i|x^{i-1}), \quad \text{where} \quad q_i(x_i|x^{i-1}) = \frac{q^i(x^i)}{q^{i-1}(x^{i-1})}.$$

In the display above, $q^i(x^i)$ denotes the marginal distribution assigned to x^i by q^n: that is, by summing over all possible x_{i+1}^n values. Now, observe that

$$q_i(x_i|x^{i-1}) = \frac{\int_\Theta \pi(\theta)p_\theta(x^{i-1})p_\theta(x_i|x^{i-1})d\theta}{\int_\Theta \pi(\theta')p_{\theta'}(x^{i-1})d\theta'} = \int_\Theta \frac{\pi(\theta)p_\theta(x^{i-1})}{\int_\Theta \pi(\theta')p_{\theta'}(x^{i-1})d\theta'}p_\theta(x_i|x^{i-1})d\theta$$
$$\int_\Theta \pi(\theta|x^{i-1})p_\theta(x_i|x^{i-1})d\theta.$$

In many cases, this predictive distribution of $q_i(\cdot|x^{i-1})$ can be computed in closed form. Furthermore, we can interpret the posterior distribution as follows:

$$\pi(\theta|x^{i-1})p_\theta(x_i|x^{i-1}) \propto \pi(\theta)p_\theta(x^{i-1})p_\theta(x_i|x^{i-1}) = \pi(\theta)\exp\left(-\log(1/p_\theta(x^{i-1}))\right)p_\theta(x_i|x^{i-1}).$$

Thus, the mixture prediction strategy is an instance of a more general class of predictors, called the *exponentially weighted* predictors.

4.1.3 Asymptotically minimax optimal redundancy with Jeffreys prior

In general, computing the capacity achieving input distribution is non-trivial, even for simple classes such as class of i.i.d. Bernoulli distributions. Instead, we now relax the requirements, instead look to find the distribution achieving asymptotically minimax optimal redundancy against the class of all i.i.d. distributions.

Theorem 123. *Let $\mathcal{X}$ denote a finite alphabet, and for any $n \geq 1$, let $\mathcal{P}_n$ denote the class of all i.i.d. distributions on $\mathcal{X}^n$. Then, Q_π where $\pi \equiv \pi_J$ is Jeffreys prior achieves asymptotically minimax optimal redundancy.*

In other words, the redundancy of any mixture π satisfies

$$R(Q_\pi, P_\theta) = \frac{m-1}{2}\log(n/2\pi e) + \frac{1}{2}\log\left(\frac{\pi(\theta)}{\sqrt{detJ(\theta)}}\right) + o(1).$$

The only π-dependent term is minimized by setting $\pi(\theta) \propto \sqrt{det(J(\theta))}$, which gives us the required result.

Remark 124. For the case considered above, Jeffreys prior is defined as

$$\pi(\theta) = c_m \frac{1}{\sqrt{\prod_{x\in\mathcal{X}}\theta(x)}}, \quad \text{with} \quad c_m \coloneqq \frac{\Gamma(m/2)}{\Gamma(1/2)^m}.$$

It is easy to verify that in this case, the predictive distribution is the "add-1/2" estimators:

$$q_i(x|x^{i-1}) = \frac{1/2 + \sum_{j=1}^{i-1}\mathbf{1}_{x_j=x}}{m/2+i-1}.$$

Thus, this prediction scheme can be combined with the arithmetic coding scheme, to get a computationally efficient way of encoding any i.i.d. source with near-optimal redundancy of $\frac{m-1}{2}\log(n/2\pi e)$.

Remark 125. This result provides a new interpretation of Jeffreys prior. In statistics, this prior distribution is used as it is invariant to a coordinate transformation (under some mild regularity assumptions).

In proving Theorem 123, we can show that the mutual information between $U \sim \pi$ and the observations can be written as

$$I(U;X^n) = \frac{m-1}{2}\log\left(\frac{n}{2\pi e}\right) + \int_\Theta \log\left(\frac{\sqrt{\det(J(u)}}{\pi(u)}\right)\pi(u)du + o(1).$$

Then, by choosing $\pi_J(u) \propto \sqrt{\det(J(u))}$, we can show that

$$\int_\Theta \pi(u)\log\left(\frac{\sqrt{\det(J(u)}}{\pi(u)}\right)\pi(u)du = -D_{\mathrm{kl}}\left(\pi \parallel \pi_J\right) + \log\left(\int_\Theta \sqrt{\det(J(u))}du\right).$$

Thus, Jeffreys prior asymptotically maximizes the mutual information between the parameter U and the observations X^n: such a prior is called a reference prior.

The proof of this statement is quite technical; instead of describing all the details, we present an outline.

4.2 Universal Compression II: Individual Sequence

We now consider the second framework of universal compression. In this setting, instead of looking at the average suboptimality, our goal is to minimize the worst case (over all sequences) suboptimality. In particular, the regret incurred by a predictor q^n, with respect to another predictor p^n, on a sequence of symbols x^n is defined as

$$\mathrm{Reg}(q^n, p^n, x^n) = \log\left(\frac{1}{q^n(x^n)}\right) - \log\left(\frac{1}{p^n(x^n)}\right).$$

Our goal is to control the worst case regret over all sequences:

$$\mathrm{Reg}(q^n, p^n) = \sup_{x^n \in \mathcal{X}^n} \mathrm{Reg}(q^n, p^{,} x^n).$$

Note that in order to achieve sublinear in n regret, we must restrict the class of distributions in which p^n can lie. Otherwise, for any q^n, there exists a sequence x^n, and a distribution p^n, such that

$$\mathrm{Reg}(q^n, p^n, x^n) \geq n \log m.$$

Thus, we are interested in the worst case regret with respect to a class of predictors/distributions $\mathcal{P}$, defined as

$$\mathrm{Reg}(q^n, \mathcal{P}) = \sup_{p^n \in \mathcal{P}} \sup_{x^n \in \mathcal{X}^n} \log\left(\frac{p^n(x^n)}{q^n(x^n)}\right).$$

4.2.1 Regret-Complexity

For a given class of distributions $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, for some parameter space Θ, define the complexity of this class of distributions as

$$\mathrm{Comp}(\Theta) := \log\left(\sum_{x \in \mathcal{X}} \sup_{\theta \in \Theta} p_\theta(x)\right).$$

We can now state the main result of this section, which characterizes the minimax regret for this class of distributions.

Theorem 126. *The minimax regret for the class of predictors $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ is equal to*

$$\mathit{Reg}(\Theta) := \inf_q \sup_\theta \sup_{x \in \mathcal{X}} \log\left(\frac{p_\theta(x)}{q(x)}\right) = \mathit{Comp}(\Theta).$$

Furthermore, the distribution that achieves this is the Normalized Maximum Likelihood (NML) distribution, defined as

$$q(x) \propto \sup_{\theta \in \Theta} p_\theta(x).$$

This is also called the Shtarkov distribution.

Proof. We prove this result in two steps:

- The first step is to show that the NML strategy has a constant regret on every outcome x.

- Next, we show that for any other prediction strategy, the regret is at least as large as the constant from the previous step.

First, we observe that for the NML strategy q^*, we have

$$\begin{aligned}\mathrm{Reg}(q^*, \Theta) &= \sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}} \log\left(\frac{p_\theta(x)}{q^*(x)}\right) = \sup_{x\in\mathcal{X}}\log\left(\frac{\sup_{\theta\in\Theta} p_\theta(x)}{q^*(x)}\right) \\ &= \sup_{x\in\mathcal{X}}\log\left(\frac{\sup_{\theta\in\Theta} p_\theta(x)}{\sup_{\theta'} p_{\theta'}(x)/\sum_{x'}\sup_{\theta'} p_{\theta'}(x')}\right) = \mathrm{Comp}(\Theta).\end{aligned}$$

Next, we consider any other predictor q, and observe that

$$\mathrm{Reg}(q, \Theta) = \sup_x \sup_\theta \log\left(\frac{p_\theta(x)}{q(x)}\right) = \sup_x \log\left(\frac{\sup_\theta p_\theta(x)}{q(x)}\right) = \sup_x \log\left(\frac{q^*(x)}{q(x)}\right) + \mathrm{Comp}(\Theta)$$

Now, we use the fact that the maximum over x is lower bounded by the average to get

$$\mathrm{Reg}(q, \Theta) \geq \sum_{x\in\mathcal{X}} q^*(x)\log\left(\frac{q^*(x)}{q(x)}\right) + \mathrm{Comp}(\Theta) = D_{\mathrm{kl}}(q^* \parallel q) + \mathrm{Comp}(\Theta).$$

□

Remark 127. When working with observations in a product space $\mathcal{X}^n$, the above statement has simple analog:

$$q^n(x^n) \propto \sup_{\theta\in\Theta} p_\theta^n(x^n),$$

where $\{p_\theta^n : \theta \in \Theta\}$ denotes a class of distributions over the product space (not necessary independent or i.i.d. distributions). This approach suffers from two drawbacks:

- in general computing the NML distribution incurs exponential complexity, which is infeasible for all but very small n and m.
- the above scheme is not 'anytime': we cannot update the NML distribution incrementally, and we must recompute it from scratch for every new observation.

Approximation of the complexity parameter for i.i.d. Bernoulli sources. For i.i.d. Bernoulli sources, we can obtain a tight approximation of the complexity using approximations of the binomial coefficients. In particular, we have $\mathcal{X} = \{0, 1\}$, $\Theta = [0, 1]$, and $\mathcal{P}_n$ denotes the class of all i.i.d. Bernoulli distributions with parameter in Θ. First, note that for any $x^n \in \mathcal{X}^n$, we have

$$p_\theta(x^n) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n 1-x_i} = \exp\left(n(D_{\mathrm{kl}}(\widehat{\theta} \parallel \theta) + h_2(\widehat{\theta}))\right), \quad \text{with} \quad \widehat{\theta} = \frac{\sum_{i=1}^n x_i}{n}.$$

This implies that

$$\sup_{\theta\in\Theta} p_\theta^n(x^n) = \exp\left(nh_2(\widehat{\theta})\right).$$

Now, we use the above expression to write

$$\begin{aligned}\exp\left(\mathrm{Comp}(\Theta)\right) &= \sum_{x^n\in\mathcal{X}^n}\sup_{\theta\in\Theta} p_\theta^n(x^n) = \sum_{x^n\in\mathcal{X}^n}\exp\left(nh_2(\widehat{\theta})\right) \\ &= \sum_{i=0}^n \binom{n}{i}\exp\left(h_2(i/n)\right).\end{aligned}$$

Now, we use the fact that for any $i \in \{1, \ldots, n-1\}$, we have

$$\binom{n}{i} = \frac{\exp\left(nh_2(i/n)\right)}{\sqrt{cn(i/n)(1-i/n)}}, \quad \text{with} \quad \frac{1}{c} \in \left[\frac{1}{8}, \frac{1}{\pi}\right].$$

Thus, using this, we get

$$\begin{aligned}\exp\left(\mathrm{Comp}(\Theta)\right) &= 2 + \sum_{i=1}^{n-1} \frac{1}{\sqrt{cn(i/n)(1-i/n)}} \;=\; 2 + \sqrt{\frac{n}{c}} \sum_{i=1}^{n-1} \frac{1}{\sqrt{(i/n)(1-i/n)}} \frac{1}{n} \\ &\approx 2 + \sqrt{\frac{n}{c}} \int_{\Theta} \frac{d\theta}{\sqrt{\theta(1-\theta)}} = 2 + \sqrt{\frac{n}{c}} \int_{\Theta} \sqrt{J_\theta} d\theta.\end{aligned}$$

Thus, on taking the logarithm, we get

$$\mathrm{Comp}(\Theta) \approx \frac{1}{2}\log\left(\frac{n}{\pi}\right) + \log\left(\int_{\Theta} \sqrt{J_\theta} d\theta\right) + O(1).$$

Thus, for the case of i.i.d. Bernoulli class, the complexity (hence minimax regret) is with a constant term of the optimal minimax redundancy.

4.3 Universal Portfolios

We study the individual sequence version of universal portfolio optimization. In particular, our goal is to design strategies that work nearly as well as the best constantly rebalanced portfolio in hindsight. More specifically, let $(\boldsymbol{b}_i)_{i\geq 1}$ denote any predictable portfolio strategy, and let $W_n(x^n)$ denote the corresponding wealth after n rounds. Our goal is to understand the value of V_n below, and find predictable investment strategies that come close to this optimum value.

$$V_n \coloneqq \max_{(\boldsymbol{b}_i)_{i\geq 1}} \min_{x^n} \frac{W_n(x^n)}{\max_{\boldsymbol{b}} W_n(\boldsymbol{b}, x^n)} = \max_{(\boldsymbol{b}_i)_{i\geq 1}} \min_{x^n} \frac{W_n(x^n)}{W_n^*(x^n)}.$$

We present two results regarding this quantity. For simplicity, we will focus on the case of two stocks (i.e., $m = 2$).

Proposition 128. *The ratio of the wealth achieved by the best predictable strategy, and the best constantly rebalanced portfolio in hindsight, is*

$$\frac{1}{V_n} = \sum_{i=0}^{n} \binom{n}{i} 2^{-nh_2(i/n)}.$$

Using Stirlings approximation, we have

$$\frac{1}{2\sqrt{n+1}} \leq V_n \leq \frac{2}{\sqrt{n}}, \quad \textit{for all } n \geq 1.$$

This implies that there exists an investment strategy that satisfies

$$\frac{W_n(x^n)}{W_n^*(x^n)} \geq \frac{1}{2\sqrt{n+1}}, \quad \textit{for all } x^n, \textit{ and } n \geq 1.$$

The betting strategy that achieves the optimal value of V_n in the statement above is not ‘anytime’, as it is the portfolio analog of Shtarkov’s distribution. We now show that a mixture strategy can nearly achieve this optimum performance.

Let μ denote any distribution over the space of probability distributions Δ_1, that characterizes all constantly rebalanced strategies for 2 stocks. Then, the mixture strategy selects $\boldsymbol{b}_{i+1}$ is defined as follows:

$$\boldsymbol{b}_{i+1} = \frac{\int \boldsymbol{b} W_i(\boldsymbol{b}, x^i) d\mu(\boldsymbol{b})}{\int W_i(\boldsymbol{b}, x^i) d\mu(\boldsymbol{b})}.$$

It is easy to verify that

$$\langle \boldsymbol{b}_{i+1}, x_{i+1} \rangle = \frac{\int \langle \boldsymbol{b}, x_{i+1} \rangle W_i(\boldsymbol{b}, x^i) d\mu(\boldsymbol{b})}{\int W_i(\boldsymbol{b}, x^i) d\mu(\boldsymbol{b})} = \frac{\int W_{i+1}(\boldsymbol{b}, x^{i+1}) d\mu(\boldsymbol{b})}{\int W_i(\boldsymbol{b}, x^i) d\mu(\boldsymbol{b})}.$$

This means that under this strategy, the wealth after i rounds is the mixture (weighted according to μ) of the wealth achieved by all constantly rebalanced portfolios.

Proposition 129. *The wealth of the mixture investment strategy on some sequence x^n, denoted by $W_n(x^n)$, satisfies the following:*

$$\frac{W_n(x^n)}{W_n^*(x^n)} \geq \frac{V_n}{\sqrt{2\pi}} \geq \frac{1}{2\sqrt{2\pi(n+1)}}.$$

Thus the performance of the mixture betting strategy is within a constant factor of the best (worst-case) performance achievable by any predictable betting strategy.

Proof. We proceed in the following steps:

Transform wealth to sum of products. Consider any investment strategy that bets $(\boldsymbol{b}_i)_{i\geq 1}$. Then, its wealth after n rounds is

$$\begin{aligned} W_n(\boldsymbol{b}^n, x^n) &= \prod_{i=1}^n (\boldsymbol{b}_{i,1}x_{i,1} + \boldsymbol{b}_{i,2}x_{i,2}) = \sum_{j^n \in \{1,2\}^n} \prod_{i=1}^n \boldsymbol{b}_{i,j_i} x_{i,j_i} \\ &:= \sum_{j^n \in \{1,2\}^n} w(j^n)x(j^n). \end{aligned}$$

In particular, for the mixture strategy we have

$$\begin{aligned} W_n(\mu, x^n) &= \int \prod_{i=1}^n (\boldsymbol{b}_1 x_{i,1} + \boldsymbol{b}_2 x_{i,2}) d\mu(\boldsymbol{b})0 = \int \sum_{j^n} \prod_{i=1}^n \boldsymbol{b}_{j_i} x_{i,j_i} d\mu(\boldsymbol{b}) \\ &= \sum_{j^n} \left(\int \prod_{i=1}^n \boldsymbol{b}_{j_i} d\mu(\boldsymbol{b}) \right) \prod_{i=1}^n x_{j_i} := \sum_{j^n} w_\mu(j^n)x(j^n). \end{aligned}$$

Step 2: Get a lower bound on the ratio. Now, for the sequence x^n, the ratio of the wealth of the mixture method, and the best constantly rebalanced portfolio in hindsight, is

$$\frac{W_n(\mu, x^n)}{W_n^*(x^n)} = \frac{\sum_{j^n} w_\mu(j^n)x(j^n)}{\sum_{j^n} w^*(j^n)x(j^n)} \geq \min_{j^n} \frac{w_\mu(j^n)}{w^*(j^n)} = \min_{j^n} \frac{\int \prod_{i=1}^n \boldsymbol{b}_{j_i} d\mu(\boldsymbol{b})}{\prod_{i=1}^n \boldsymbol{b}^*_{j_i}}. \tag{4.3}$$

This shows that the worst case stock market is a horse race!

Step 3: Calculate the denominator in (4.3). Let us assume that the $j^n \in \{1,2\}^n$ that achieves the minimum in (4.3) has l $1's$. Then, the optimal $\boldsymbol{b}^*$ for this sequence is $(b, 1-b)$, where $b = l/n$. Furthermore, the denominator in this case is $2^{-nh_2(l/n)}$.

Step 4: Calculate the numerator. If we take the Jeffreys prior, then we have $\mu(b) = 1/(\pi\sqrt{b(1-b)})$. This implies that the numerator in (4.3) is

$$\begin{aligned} \int \prod_{i=1}^n b_{j_i} d\mu(b) &= \int b^l (1-b)^{n-l} \frac{1}{\pi\sqrt{b(1-b)}} db = \frac{1}{\pi} \int b^{l+1/2-1}(1-b)^{n-l+1/2-1} db \\ &= \frac{\Gamma(l+1/2)\Gamma(n-l+1/2)}{\Gamma(n+1)}. \end{aligned} \tag{4.4}$$

Plugging (4.4) into (4.3) gives us

$$\frac{W_n(\mu, x^n)}{W_n^*(x^n)} \geq \frac{\Gamma(l+1/2)\Gamma(n-l+1/2)}{\Gamma(n+1)2^{-nh_2(l/n)}} \geq \frac{V_n}{\sqrt{2\pi}} \geq \frac{1}{2\sqrt{2\pi(n+1)}}.$$

Or equivalently, we have

$$\log\left(W_n^*(x^n)\right) - \log\left(W_n(\mu, x^n)\right) \leq \frac{1}{2}\log\left(8\pi(n+1)\right).$$

□

4.4 Coin Betting

In this section, we study a special case of the general case of the portfolio optimization problem.

Definition 130. Let $W_0 = 1$, and proceed as follows, for $t = 1, 2, \ldots$:

- Choose a $\lambda_t \in [-1, 1]$.
- Observe the next value $c_t \in [-1, 1]$.
- Update the wealth $W_t = W_{t-1} \times (1 + \lambda_t c_t)$.

The goal is to minimize the following regret, with respect to the best constant bet in hindsight:

$$\mathcal{R}_n = \sup_{\lambda \in [-1,1]} \sum_{t=1}^{n} \log(1 + \lambda c_t) - \sum_{t=1}^{n} \log(1 + \lambda_t c_t).$$

The first observation is that this is simply a portfolio optimization problem with two stocks, whose price relatives are $1 + c_t$ and $1 - c_t$ respectively in round t. Then, any portfolio allocation $\boldsymbol{b} = (b, 1 - b)$ on these two stocks corresponds to betting $\lambda_b = 2b - 1$ in the coin betting game:

$$\langle \boldsymbol{b}, \boldsymbol{x} \rangle = b(1 + c_t) + (1 - b)(1 - c_t) = 1 + (2b - 1)c_t = 1 + \lambda_b c_t.$$

For this problem, we can define the simple mixture strategy that selects λ_t as

$$\lambda_t = \frac{\int_{-1}^{1} \lambda \, W_{t-1}^{\lambda}(c^{t-1}) d\mu(\lambda)}{\int_{-1}^{1} W_{t-1}^{\lambda}(c^{t-1}) d\mu(\lambda)}, \quad \text{where} \quad W_{t-1}^{\lambda}(c^{t-1}) = \prod_{i=1}^{t-1} (1 + \lambda c_t),$$

where μ is the Beta $(1/2)$ distribution.

Application: Online linear optimization. TODO

Part II

Applications

Chapter 5

Application I: From Universal Compression to Sequential Inference

5.1 Testing by Betting

5.2 Hypothesis Testing

5.3 Confidence Sequences

Bibliography

I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

J. L. Kelly. A new interpretation of information rate. *the bell system technical journal*, 35(4):917–926, 1956.

Y. Peres. Iterating von neumann's procedure for extracting random bits. *The Annals of Statistics*, pages 590–597, 1992.

Y. Polyanskiy and Y. Wu. *Information theory: from coding to learning.* Cambridge University Press, 2023.