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

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

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.

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

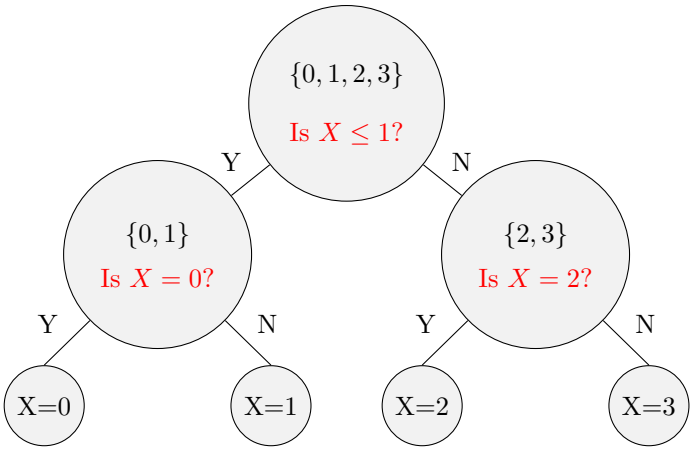

Figure 1.1: The figure shows the decision tree for the optimal strategy (i.e., binary search) for the guessing game with X drawn uniformly at random from $\mathcal{X} = \{0, 1, 2, 3\}$. Each query is chosen to ensures that the outcomes (i.e., Y or N) are equiprobable. In other words, this strategy proceeds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

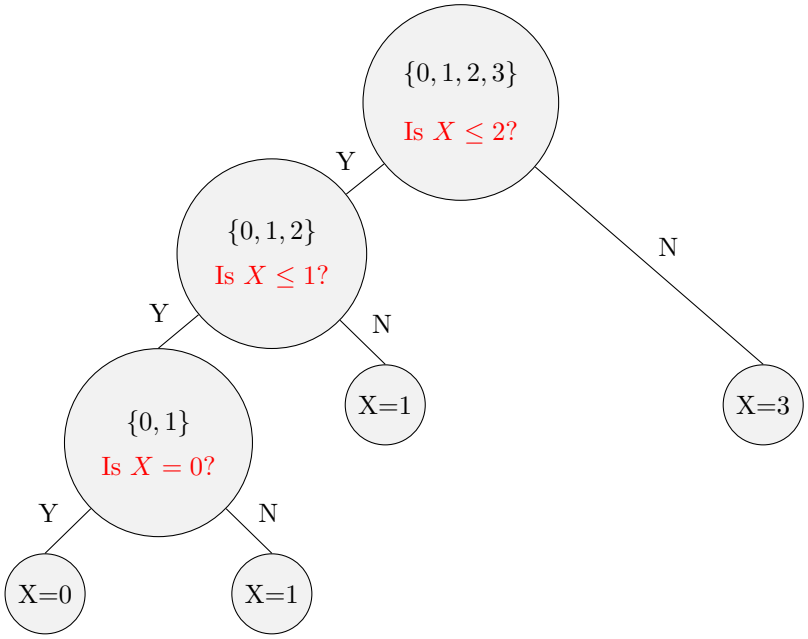

Figure 1.2: The figure shows the decision tree for the optimal strategy for a non-uniform probability distribution over $\mathcal{X}$. The key observation is that this strategy assigns fewer questions (or a shorter codeword) to the higher probability symbol (3), and more questions to the less probable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

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

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

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

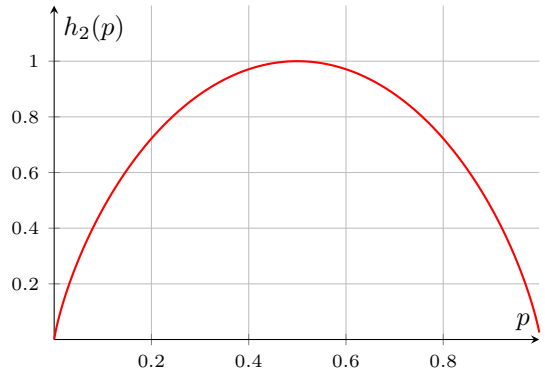

Figure 1.3: Variation of the binary entropy $h_2(p) = -p\log(p) - \bar{p}\log(\bar{p})$ with the parameter p. Note that $h_2(p)$ is zero for $p \in \{0, 1\}$, when the random variable is deterministic, and is maximized at $p = 0.5$, which corresponds to the maximum uncertainty. Also note that the qualitative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

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

□

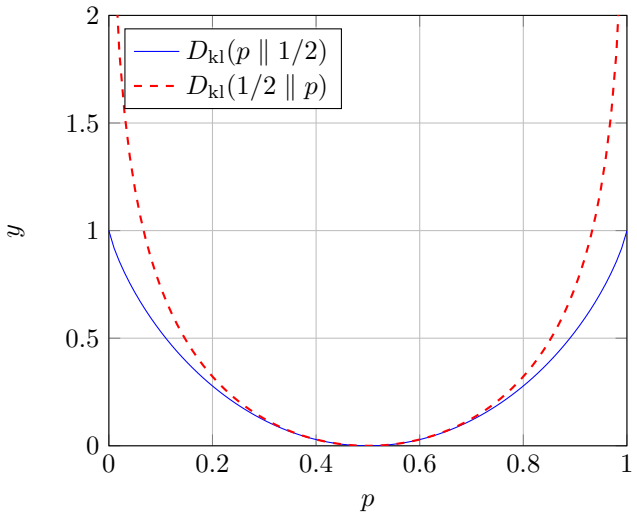

Figure 1.4: Plots of $D_{\mathrm{kl}}(P_X \parallel P_Y)$ and $D_{\mathrm{kl}}(P_Y \parallel P_X)$, where $X \sim \text{Bernoulli}(p)$ and $X \sim \text{Bernoulli}(1/2)$, as p varies in $[0, 1]$. The figures illustrates the locally quadratic behavior of relative entropy, and also shows that it is not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

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

□

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 bet 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 $E = \{(X_1, X_2) = (0,1) \text{ or } (X_1, X_2) = (1,0)\}$.
- 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

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

  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

$$D_{kl}\left(P_{X^n} \parallel Q_{X^n}\right) = D_{kl}\left(P_{X^n} \parallel \prod_{i=1}^{n} P_{X_i}\right) + \sum_{i=1}^{n} D_{kl}\left(P_{X_i} \parallel Q_{X_i}\right) \tag{1.3}$$

$$\overset{(a)}{\geq} \sum_{i=1}^{n} D_{kl}\left(P_{X_i} \parallel Q_{X_i}\right). \tag{1.4}$$

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

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

Proof. Introduce the notations $X^i = (X_1, \dots, X_i)$ and $X_i^j = (X_i, \dots, X_j)$ 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 $X^n = (X_1, \dots, X_n)$ 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 $(x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)$: 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 $\lambda \notin \{0, 1\}$. 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 $X \to Y \to Z$ for a Markov chain; that is, $X \perp Z|Y$. Then, we have*

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

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 $H(X|Z) \le H(X|Z,Y)$. Furthermore, due to the Markov property, $X \perp Z|Y$, which means that $H(X|Z,Y) = H(X|Y)$. This implies that

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

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*

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

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

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

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*

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

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

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

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,

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

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:

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

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 $Z \sim$ 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

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

□

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

$$h(X) = -\int_{\mathcal{X}} f_X(x) \log(f_X(x)) dx.$$

Similarly, conditional entropy of Y given X is defined as

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

Example 37. Suppose $\mathcal{X} = [0, a]$, and X is a the uniformly distributed random variable over $\mathcal{X}$ with density $f_X(x) = 1/a$ for all $x \in \mathcal{X}$. Then, the differential entropy of X is equal to

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

Thus, $h(X) = 0$ for $a = 1$, $h(X) < 0$ for $a < 1$, and $h(X) > 0$ for $a > 1$. 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. 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

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

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.

An interesting corollary of this result is the co

Definition 39. 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

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

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

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

Example 40. 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, for two distributions with equal variance, the relative entropy is proportional to the difference in mean squared.

Example 41. 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, and Q has a Gaussian distribution. Clearly, $P \ll Q$ and $Q \ll P$. However,

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

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 42. *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

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

□

Proposition 43. *Let $\mathcal{P}_a$ denote all continuous distributions supported on $[0, a]$. Then, the distribution from $\mathcal{P}_a$ with the largest entropy is the uniform distribution.*

Proof. □

Definition 44. The mutual information between two continuous random variables X and Y is defined as

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

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 45. 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 $\{i\Delta : 1 \leq i \leq a/\Delta\}$. 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 46. 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

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

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 47. *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 46).

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 48. 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 49. 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,

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

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 $\Delta \to 0$.

2.2 General Distributions*

Fact 50. *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*

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

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

Chain rule for Radon-nikodym derivatives.

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 51. 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

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

Fact 52. *An equivalent definition of the relative entropy between P and Q, is*

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

See Lemma 2.4 of Polyanskiy and Wu for a proof.

Definition 53. 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 54. 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 55 (Disintegration Theorem)**.** *Suppose P_{XY} is a joint distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{Y}$ is standard Borel. Then, there exists a Markov kernel K, such that for any measurable $E \subset \mathcal{X} \times \mathcal{Y}$, we have*

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

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

Definition 56. 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 57. Having defined relative entropy, and conditional relative entropy for general distributions, we can immediately use them to define mutual information and conditional mutual information.

Hence, 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). From this definition, we can then obtain the definitions of differential entropy, and mutual information.

2.2.1 Variational Definition I: Gelfand-Yaglom-Perez

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 58. 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:

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

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 59. *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 $\le$ and $\ge$ 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

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

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 $D_{\mathrm{kl}}(P \parallel Q) = \lim_{c \to \infty} \mathbb{E}_Q[\varphi(X)\mathbf{1}_{X \le c}]$ 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 $\mathcal{E} = \{E_0, \ldots, E_n\}$, where the sets E_0 through E_n are defined as

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

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

$$\mathbb{E}_Q[\varphi(X_c)] - \delta \leq \mathbb{E}_Q[\varphi(Y_n)] = \sum_{j=0}^{n-1} Q(E_j)\varphi(j\epsilon). \tag{2.1}$$

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 (2.1), we get

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

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 60. *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 61. *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. As in the previous case, we prove this result in two steps to show that both the $\leq$ and $\geq$ hold.

First, we show that $D_{\mathrm{kl}}(P \parallel Q)$ is an upper bound on the supremum. Without loss of generality, we assume that $P \ll Q$ and $D_{\mathrm{kl}}(P \parallel Q) < \infty$. 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

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

Now, note that we also have $Q \ll Q^f$, which implies that $P \ll Q^f$. Hence, by the chain rule for Radon-Nikodym derivatives, we have $dQ^f/dQ = dQ^f/dP \times dP/dQ$, which leads to

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

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 59. 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

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

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 59. □

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 62. *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 $S = (Z_1, \dots, Z_n) \in \mathcal{Z}^n$ consisting on 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:

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

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

Proposition 63 (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*

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

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

$$I(S;H) = \sup_f \mathbb{E}_{P_{SH}}[f(S,H)] - \log \mathbb{E}_{P_S \times P_H}\left[e^{f(S,H)}\right]. \tag{2.2}$$

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

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

For any $\lambda \in \mathbb{R}$, plugging λf in (2.2), we get

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

On optimizing for λ, we get

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

On rearranging, we get the required

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

□

2.3 f-divergences

Definition 64. 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

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

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

Remark 65. 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. Here, we state an analog of Theorem 61 for f-divergences on $\mathcal{X} = \mathbb{R}^d$.

Theorem 66. *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

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

□

Application. We will use f-divergences for constructing sequential tests in the second part of the course.

Chapter 3

Compression and Gambling

- Definition of a D-ary code
- Singular, Uniquely Decodable, Prefix Codes
- Krafts inequality for prefix codes, and UD codes
- Entropy as a lower bound on the codeword length
- Relative entropy as a measure of redundancy
- Achievability: Shannon Codes, SFE codes, Huffman code
- Doubling rate in gambling
- Value of side-information in gambling
- Portfolio optimization + Proportion (or Kelly) Betting

3.1 Source Codes

Definition 67. 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 $C(x^k) = C(x_1)C(x_2)\ldots C(x_k)$.

Definition 68. 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 $C(x')$.

Remark 69. Nonsingular codes cannot be applied in general to streams of symbols. Uniquely decodable 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

In this section, we will obtain a correspondence between probability distributions and compression schemes. In particular, we show that every prefix code (in fact, every UD code) induces a probability distribution on the alphabet.

Theorem 70. *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.

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, given $(l_i)_{i\geq 1}$ satisfying Kraft's inequality, we construct codewords as the binary representation of the lower endpoints of the following disjoint intervals:

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

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 71. *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 a 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 $\mathcal{X}_N^m$ 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 $\mathcal{X}_N^*$. This implies that $a(m)2^{-m} \leq 1$, and thus

$$\begin{aligned}&\left(\sum_{x \in \mathcal{X}_N} 2^{-l(x)}\right)^k \leq k l_{\max} \\ \implies &\sum_{x \in \mathcal{X}_N} 2^{-l(x)} \leq k^{1/k} l_{\max}^{1/k}.\end{aligned}$$

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

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

Since N was arbitrary, we can then conclude that

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

□

3.3 Lower Bound and SFE coding

Entropy as a lower bound on average codeword length. We now observe that for a random variable $X \sim P_X$ taking values on the alphabet $\mathcal{X}$, the average codeword length for any uniquely decodable code C is lower bounded by the entropy of X.

Theorem 72. *Suppose $X \sim P_X$ is a random variable taking values in a countable alphabet $\mathcal{X}$. If C is any uniquely decodable coding scheme for X, then, we have*

$$L_C(X) = \sum_{x \in \mathcal{X}} p(x) l(x) \geq H(X) = \sum_{x \in \mathcal{X}} p(x) \log(1/p(x)).$$

Proof. Since C is uniquely decodable, it satisfies Kraft's inequality. Thus, the term $A = \sum_{x \in \mathcal{X}} 2^{-l(x)}$ 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:

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

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

The proof of the above result also tells us when the optimal bound is achieved: if $A = 1$, and $P_X = R$.

Typical Set Coding. We saw an example of a lossless compression scheme which achieves close to the optimal compression rate in Remark 9. However, that scheme is mainly a theoretical construction, and is not feasible for practical compression tasks due to its exponential complexity.

Shannon Fano Elias Coding. We now present a practical compression scheme that achieves the optimal compression rate for large block lengths.

3.4 Gambling, Doubling Rate

3.5 Portfolio Optimization + Kelly Betting

Chapter 4

Universal Compression and Gambling

4.1 Universal Compression

4.1.1 Redundancy-Capacity

4.1.2 Regret-Complexity

4.1.3 Mixture Method

4.2 Universal Portfolios

Part II

Applications

Chapter 5

Application I: From Universal Compression to Sequential Inference

Bibliography

I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

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.