Chapter 1

Review of basic (and not so basic) concepts in information theory

Readings covering the material in this set of notes: Chapter 2 of Cover and Thomas [2], which covers all of the results for the discrete case. For the more advanced (measure-theoretic) version, see Chapter 5 of Gray [4] (available on Bob Gray's webpage), or Chapter 7 of the second edition of the same book.

1.1 Basics of Information Theory

In this section, we review the basic definitions in information theory, including (Shannon) entropy, KL-divergence, mutual information, and their conditional versions. Before beginning, I must make an apology to any information theorist reading these notes: any time we use a log, it will always be base-e. This is more convenient for our analyses, and it also (later) makes taking derivatives much nicer.

In this first section, we will assume that all distributions are discrete; this makes the quantities somewhat easier to manipulate and allows us to completely avoid any complicated measure-theoretic quantities. In Section 1.2 of this note, we show how to extend the important definitions (for our purposes)—those of KL-divergence and mutual information—to general distributions, where basic ideas such as entropy no longer make sense. However, even in this general setting, we will see we essentially lose no generality by assuming all variables are discrete.

1.1.1 Definitions

Entropy: We begin with a central concept in information theory: the entropy. Let P be a distribution on a finite (or countable) set $\mathcal{X}$, and let p denote the probability mass function associated with P. That is, if X is a random variable distributed according to P, then $P(X = x) = p(x)$. The *entropy of* X (or of P) is defined as

$$H(X) := -\sum_x p(x) \log p(x).$$

Because $p(x) \leq 1$ for all x, it is clear that this quantity is positive. We will show later that if $\mathcal{X}$ is finite, the maximum entropy distribution on $\mathcal{X}$ is the uniform distribution, setting $p(x) = 1/|\mathcal{X}|$ for all x, which has entropy $\log(|\mathcal{X}|)$.

While we do not explore it in this class, there is an operational interpretation of entropy via Shannon's source-coding theorem (see, for example, Chapter 5 of Cover and Thomas [2]). In particular, Shannon's source coding theorem states that if we wish to encode a random variable X, distributed according to P, with a k-ary string (i.e. each entry of the string takes on one of k values), then the minimal expected length of the encoding is given by $H(X) = -\sum_x p(x) \log_k p(x)$. Moreover, this is achievable (to within a length of at most 1 symbol) by using Huffman codes (among many other types of codes). As an example of this interpretation, we may consider encoding a random variable X with equi-probable distribution on m items, which has $H(X) = \log(m)$. In base-2, this makes sense: we simply assign an integer to each item and encode each integer with the natural (binary) integer encoding of length $\lceil \log m \rceil$.

We can also define the *conditional entropy*, which is the amount of information left in a random variable after observing another. In particular, we define

$$H(X \mid Y = y) = -\sum_x p(x \mid y) \log p(x \mid y) \quad \text{and} \quad H(X \mid Y) = \sum_y p(y) H(X \mid Y = y),$$

where $p(x \mid y)$ is the p.m.f. of X given that $Y = y$.

KL-divergence: Now we define two additional quantities, which are actually *much more* fundamental than entropy: they can always be defined for any distributions and any random variables, as they measure distance between distributions. Entropy simply makes no sense for non-discrete random variables, let alone random variables with continuous and discrete components, though it proves useful for some of our arguments and interpretations.

Before defining these quantities, we recall the definition of a convex function $f : \mathbb{R}^k \to \mathbb{R}$ as any bowl-shaped function, that is, one satisfying

$$f(\lambda x + (1 - \lambda) y) \le \lambda f(x) + (1 - \lambda) f(y) \tag{1.1.1}$$

for all $\lambda \in [0, 1]$, all x, y. The function f is *strictly* convex if the convexity inequality (1.1.1) is strict for $\lambda \in (0, 1)$ and $x \neq y$. We recall a standard result:

Proposition 1.1 (Jensen's inequality)**.** *Let f be convex. Then for any random variable X,*

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)].$$

Moreover, if f is strictly convex*, then $f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$ unless X is constant.*

Now we may define and provide a few properties of the KL-divergence. Let P and Q be distributions defined on a discrete set $\mathcal{X}$. The *KL-divergence* between them is

$$D_{\rm kl}\left(P \| Q\right) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

We observe immediately that $D_{\rm kl}\left(P \| Q\right) \ge 0$. To see this, we apply Jensen's inequality (Proposition 1.1) to the function $-\log$ and the random variable $q(X)/p(X)$, where X is distributed according to P:

$$\begin{aligned} D_{\rm kl}\left(P \| Q\right) &= -\mathbb{E}\left[\log \frac{q(X)}{p(X)}\right] \ge -\log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] \\ &= -\log\left(\sum_x p(x) \frac{q(x)}{p(x)}\right) = -\log(1) = 0. \end{aligned}$$

Moreover, as log is strictly convex, we have $D_{\rm kl}\left(P \| Q\right) > 0$ unless $P = Q$.

Mutual information: Having defined KL-divergence, we may now describe the information content between two random variables X and Y. The *mutual information* $I(X;Y)$ between X and Y is the KL-divergence between their joint distribution and their products (marginal) distributions. More mathematically,

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{1.1.2}$$

We can rewrite this in several ways. First, using Bayes' rule, we have $p(x,y)/p(y) = p(x \mid y)$, so

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(y)p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \\ &= -\sum_x \sum_y p(y)p(x \mid y) \log p(x) + \sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y) \\ &= H(X) - H(X \mid Y). \end{aligned}$$

Similarly, we have $I(X;Y) = H(Y) - H(Y \mid X)$, so mutual information can be thought of as the amount of entropy removed (on average) in X by observing Y. We may also think of mutual information as measuring the similarity between the joint distribution of X and Y and their distribution when they are treated as independent.

Comparing the definition (1.1.2) to that for KL-divergence, we see that if P_{XY} is the joint distribution of X and Y, while P_X and P_Y are their marginal distributions (distributions when X and Y are treated independently), then

$$I(X;Y) = D_{\rm kl}\left(P_{XY} \| P_X \times P_Y\right) \geq 0.$$

Moreover, we have $I(X;Y) > 0$ unless X and Y are independent.

As with entropy, we may also define the *conditional information between X and Y given Z*, which is the mutual information between X and Y when Z is observed (on average). That is,

$$I(X;Y \mid Z) := \sum_z I(X;Y \mid Z = z)p(z) = H(X \mid Z) - H(X \mid Y, Z) = H(Y \mid Z) - H(Y \mid X, Z).$$

1.1.2 Properties, data processing

We now illustrate several of the properties of entropy, KL divergence, and mutual information; these allow easier calculations and analysis.

Chain rules: We begin by describing relationships between collections of random variables $X_1, \ldots, X_n$ and individual members of the collection. (Throughout, we use the notation $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$ to denote the sequence of random variables from indices i through j.)

For the entropy, we have the simplest chain rule:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1^{n-1}).$$

This follows from the standard decomposition of a probability distribution $p(x,y) = p(x)p(y \mid x)$. to see the chain rule, then, note that

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x)p(y \mid x) \log p(x)p(y \mid x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(x) - \sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) = H(X) + H(Y \mid X). \end{aligned}$$

Now set $X = X_1^{n-1}$, $Y = X_n$, and simply induct.

As an immediate corollary to the chain rule for entropy, we see that mutual information also obeys a chain rule:

$$I(X; Y_1^n) = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

Indeed, we have

$$I(X; Y_1^n) = H(Y_1^n) - H(Y_1^n \mid X) = \sum_{i=1}^n \left[H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})\right] = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

The KL-divergence obeys similar chain rules.

As a corollary of the chain rule for mutual information, we obtain the well-known result that *conditioning reduces entropy*:

$$H(X \mid Y) \le H(X) \text{ because } I(X; Y) = H(X) - H(X \mid Y) \ge 0.$$

So on average, knowing about a variable Y can only decrease your uncertainty about X.

Data processing inequalities: A standard problem in information theory (and statistical inference) is to understand the degradation of a signal after it is passed through some noisy channel (or observation process). The simplest of such results, which we will use frequently, is that we can only lose information by adding noise. In particular, assume we have the Markov chain

$$X \to Y \to Z.$$

Then we obtain the classical *data processing inequality*.

Proposition 1.2. *With the above Markov chain, we have $I(X; Z) \le I(X; Y)$.*

Proof We expand the mutual information $I(X; Y, Z)$ in two ways:

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y \mid Z) \\ &= I(X; Y) + \underbrace{I(X; Z \mid Y)}_{=0}, \end{aligned}$$

where we note that the final equality follows because X is independent of Z given Y:

$$I(X; Z \mid Y) = H(X \mid Y) - H(X \mid Y, Z) = H(X \mid Y) - H(X \mid Y) = 0.$$

Since $I(X; Y \mid Z) \ge 0$, this gives the result. □

1.2 General definitions of divergence and mutual information

Having given our basic definitions of mutual information and divergence, we now show how the definitions of KL-divergence and mutual information extend to arbitrary distributions P and Q and arbitrary sets $\mathcal{X}$. This requires a bit of setup, including defining set algebras (which, we will see, simply correspond to quantization of the set $\mathcal{X}$), but allows us to define divergences in full generality.

1.2.1 Partitions, algebras, and quantizers

Let $\mathcal{X}$ be an arbitrary space. A *quantizer* on $\mathcal{X}$ is any function that maps $\mathcal{X}$ to a finite collection of integers. That is, fixing $m < \infty$, a quantizer is any function $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$. In particular, a quantizer q partitions the space $\mathcal{X}$ into the subsets of $x \in \mathcal{X}$ for which $\mathsf{q}(x) = i$. A related notion—we will see the precise relationship presently—is that of an algebra of sets on $\mathcal{X}$. We say that a collection of sets $\mathcal{A}$ is an *algebra* on $\mathcal{X}$ if the following are true:

1. The set $\mathcal{X} \in \mathcal{A}$.
2. The collection of sets $\mathcal{A}$ is closed under finite set operations: union, intersection, and complementation. That is, $A, B \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$, $A \cap B \in \mathcal{A}$, and $A \cup B \in \mathcal{A}$.

There is a 1-to-1 correspondence between quantizers—and their associated partitions of the set $\mathcal{X}$—and finite algebras on a set $\mathcal{X}$, which we discuss briefly.[1] It should be clear that there is a one-to-one correspondence between finite *partitions* of the set $\mathcal{X}$ and quantizers q, so we must argue that finite partitions of $\mathcal{X}$ are in one-to-one correspondence with finite algebras defined over $\mathcal{X}$.

In one direction, we may consider a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$. Let the sets $A_1, \ldots, A_m$ be the partition associated with q, that is, for $x \in A_i$ we have $\mathsf{q}(x) = i$, or $A_i = \mathsf{q}^{-1}(\{i\})$. Then we may define an algebra $\mathcal{A}_\mathsf{q}$ as the collection of all finite set operations performed on $A_1, \ldots, A_m$ (note that this is a finite collection, as finite set operations performed on the partition $A_1, \ldots, A_m$ induce only a finite collection of sets).

For the other direction, consider a finite algebra $\mathcal{A}$ over the set $\mathcal{X}$. We can then construct a quantizer $\mathsf{q}_\mathcal{A}$ that corresponds to this algebra. To do so, we define an *atom* of $\mathcal{A}$ as any non-empty set $A \in \mathcal{A}$ such that if $B \subset A$ and $B \in \mathcal{A}$, then $B = A$ or $B = \emptyset$. That is, the atoms of $\mathcal{A}$ are the "smallest" sets in $\mathcal{A}$. We claim there is a unique partition of $\mathcal{X}$ with atomic sets from $\mathcal{A}$; we prove this inductively.

Base case: There is at least 1 atomic set, as $\mathcal{A}$ is finite; call it A_1.

Induction step: Assume we have atomic sets $A_1, \ldots, A_k \in \mathcal{A}$. Let $B = (A_1 \cup \cdots \cup A_k)^c$ be their complement, which we assume is non-empty (otherwise we have a partition of $\mathcal{X}$ into atomic sets). The complement B is either atomic, in which case the sets $\{A_1, A_2, \ldots, A_k, B\}$ are a partition of $\mathcal{X}$ consisting of atoms of $\mathcal{A}$, or B is not atomic. If B is not atomic, consider all the sets of the form $A \cap B$ for $A \in \mathcal{A}$. Each of these belongs to $\mathcal{A}$, and at least one of them is atomic, as there is a finite number of them. This means there is a non-empty set $A_{k+1} \subset B$ such that A_{k+1} is atomic.

By repeating this induction, which must stop at some finite index m as $\mathcal{A}$ is finite, we construct a collection $A_1, \ldots, A_m$ of disjoint atomic sets in $\mathcal{A}$ for which and $\cup_i A_i = \mathcal{X}$. (The uniqueness is an exercise for the reader.) Thus we may define the quantizer $\mathsf{q}_\mathcal{A}$ via

$$\mathsf{q}_\mathcal{A}(x) = i \quad \text{when } x \in A_i.$$

1.2.2 KL-divergence

In this section, we present the general definition of a KL-divergence, which holds for *any* pair of distributions. Let P and Q be distributions on a space $\mathcal{X}$. Now, let $\mathcal{A}$ be a finite algebra on $\mathcal{X}$ (as in the previous section, this is equivalent to picking a partition of $\mathcal{X}$ and then constructing the

[1] Pedantically, this one-to-one correspondence holds up to permutations of the partition induced by the quantizer.

associated algebra), and assume that its atoms are $\mathsf{atoms}(\mathcal{A})$. The KL-divergence between P and Q *conditioned on* $\mathcal{A}$ is

$$D_{\rm kl}\left(P\|Q \mid \mathcal{A}\right) := \sum_{A \in \mathsf{atoms}(\mathcal{A})} P(A) \log \frac{P(A)}{Q(A)}.$$

That is, we simply sum over the partition of $\mathcal{X}$. Another way to write this is as follows. Let $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$ be a quantizer, and define the sets $A_i = \mathsf{q}^{-1}(\{i\})$ to be the pre-images of each i (i.e. the different quantization regions, or the partition of $\mathcal{X}$ that q induces). Then the *quantized* KL-divergence between P and Q is

$$D_{\rm kl}\left(P\|Q \mid \mathsf{q}\right) := \sum_{i=1}^{m} P(A_i) \log \frac{P(A_i)}{Q(A_i)}.$$

We may now give the fully general definition of KL-divergence: the KL-divergence between P and Q is defined as

$$\begin{aligned} D_{\rm kl}\left(P\|Q\right) &:= \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathcal{A}\right) \quad \text{such that } \mathcal{A} \text{ is a finite algebra on } \mathcal{X}\right\} \\ &= \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathsf{q}\right) \quad \text{such that } \mathsf{q} \text{ quantizes } \mathcal{X}\right\}. \end{aligned} \tag{1.2.1}$$

This also gives a rigorous definition of mutual information. Indeed, if X and Y are random variables with joint distribution P_{XY} and marginal distributions P_X and P_Y, we simply define

$$I(X;Y) = D_{\rm kl}\left(P_{XY}\|P_X \times P_Y\right).$$

When P and Q have densities p and q, the definition (1.2.1) reduces to

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx,$$

while if P and Q both have probability mass functions p and q, then—as we will see in the homework—the definition (1.2.1) is equivalent to

$$D_{\rm kl}\left(P\|Q\right) = \sum_{x} p(x) \log \frac{p(x)}{q(x)},$$

precisely as in the discrete case.

We remark in passing that if the set $\mathcal{X}$ is a product space, meaning that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ for some $n < \infty$ (this is the case for mutual information, for example), then we may assume our quantizer *always* quantizes sets of the form $A = A_1 \times A_2 \times \cdots \times A_n$, that is, Cartesian products. Written differently, when we consider algebras on $\mathcal{X}$, the atoms of the algebra may be assumed to be Cartesian products of sets, and our partitions of $\mathcal{X}$ can always be taken as Cartesian products. (See Gray [4, Chapter 5].) Written slightly differently, if P and Q are distributions on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and q^i is a quantizer for the set $\mathcal{X}_i$ (inducing the partition $A_1^i, \ldots, A_{m_i}^i$ of $\mathcal{X}_i$) we may define

$$D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) = \sum_{j_1, \ldots, j_n} P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n) \log \frac{P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}{Q(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}.$$

Then the general definition (1.2.1) of KL-divergence specializes to

$$D_{\rm kl}\left(P\|Q\right) = \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) \quad \text{such that } \mathsf{q}^i \text{ quantizes } \mathcal{X}_i\right\}.$$

So we only need consider "rectangular" sets in the definitions of KL-divergence.

Measure-theoretic definition of KL-divergence If you have never seen measure theory before, skim this section; while the notation may be somewhat intimidating, it is fine to always consider only continuous or fully discrete distributions. We will describe an interpretation that will mean for our purposes that one never needs to really think about measure theoretic issues.

The general definition (1.2.1) of KL-divergence is equivalent to the following. Let μ be a measure on $\mathcal{X}$, and assume that P and Q are absolutely continuous with respect to μ, with densities p and q, respectively. (For example, take $\mu = P + Q$.) Then

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x). \tag{1.2.2}$$

The proof of this fact is somewhat involved, requiring the technology of Lebesgue integration. (See Gray [4, Chapter 5].)

For those who have not seen measure theory, the interpretation of the equality (1.2.2) should be as follows. When integrating a function $f(x)$, replace $\int f(x) d\mu(x)$ with one of two pairs of symbols: one may simply think of $d\mu(x)$ as dx, so that we are performing standard integration $\int f(x) dx$, or one should think of the integral operation $\int f(x) d\mu(x)$ as summing the argument of the integral, so $d\mu(x) = 1$ and $\int f(x) d\mu(x) = \sum_x f(x)$. (This corresponds to μ being "counting measure" on $\mathcal{X}$.)

1.2.3 f-divergences

A more general notion of divergence is the so-called f-divergence, or Ali-Silvey divergence [1, 3] (see also the alternate interpretations in the article by Liese and Vajda [5]). Here, the definition is as follows. Let P and Q be probability distributions on the set $\mathcal{X}$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. Assume w.l.o.g. that P and Q are absolutely continuous with respect to the base measure μ. The f divergence between P and Q is

$$D_f\left(P\|Q\right) := \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{1.2.3}$$

In the first homework set, you will explore several properties of f-divergences, including a quantized representation equivalent to that for the KL-divergence (1.2.1).

Examples We give three examples of f-divergences here.

1. KL-divergence: by taking $f(t) = t \log t$, which is convex and satisfies $f(1) = 0$, we obtain $D_f\left(P\|Q\right) = D_{\rm kl}\left(P\|Q\right)$.
2. KL-divergence, reversed: by taking $f(t) = -\log t$, we obtain $D_f\left(P\|Q\right) = D_{\rm kl}\left(Q\|P\right)$.
3. The *total variation distance* between probability distributions P and Q defined on a set $\mathcal{X}$ is defined as the maximum difference between probabilities they assign on subsets of $\mathcal{X}$:

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} \left|P(A) - Q(A)\right|. \tag{1.2.4}$$

   Note that (by considering compliments $P(A^c) = 1 - P(A)$) the absolute value on the right hand side is unnecessary. The total variation distance, as we shall see later in the course, is very important for verifying the optimality of different tests, and appears in the measurement of

difficulty of solving hypothesis testing problems. An important inequality, known as *Pinsker's inequality*, is that

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P \| Q\right). \tag{1.2.5}$$

By taking $f(t) = \frac{1}{2}|t-1|$, we obtain the total variation distance. Indeed, we have

$$\begin{aligned} D_f\left(P\|Q\right) &= \frac{1}{2}\int \left|\frac{p(x)}{q(x)} - 1\right| q(x) d\mu(x) = \frac{1}{2}\int |p(x) - q(x)| d\mu(x) \\ &= \frac{1}{2}\int_{x: p(x) > q(x)} \left[p(x) - q(x)\right] d\mu(x) + \frac{1}{2}\int_{x: q(x) > p(x)} \left[q(x) - p(x)\right] d\mu(x) \\ &= \frac{1}{2}\sup_{A \subset \mathcal{X}} \left[P(A) - Q(A)\right] + \frac{1}{2}\sup_{A \subset \mathcal{X}} \left[Q(A) - P(A)\right] = \|P - Q\|_{\rm TV}. \end{aligned}$$

Bibliography

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[3] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[4] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[5] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.