1 Introduction to Information Theory

We begin with an overview of probability on discrete spaces. Let the discrete set of outcomes of some experiment be Ω. The set Ω is usually called the *sample space.* Subsets of this set are called the *events.* A *probability distribution* on Ω is a function $P : \Omega \to [0, 1]$ such that the following are true.

1. $P(\emptyset) = 0$. When we do the experiment, we stipulate that some outcome in Ω should occur. Thus, the probability that no outcome in Ω occurs, should be zero.

2. We would like to state that the probability of a set S of outcomes is the sum of the probabilities of the individual outcomes in S.

   We generalize this to say the following. Suppose S is partitioned into disjoint sets $A_1, A_2, \ldots$. Then S can be viewed either as a whole, or as a union of the different A_is. Since the set is the same when viewed in both these ways, we have the stipulation that the probability of S should be the sum of the probabilities of the different A_is. We have the following condition.

   For disjoint events $A_1, A_2, \ldots,$
   $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

3. Either an event A happens, or it does not happen. Since these events are mutually exclusive and exhaustive, by the previous condition, their probabilities have to add up to the probability of the entire space of outcomes. Set $P(\Omega) = 1$. Thus we have the following condition. (Note that, the values set to Ω and $\emptyset$ satisfy the stipulation.)

   For any event A, $P(A^c) = 1 - P(A)$.

Then the pair (Ω, P) is called a *discrete probability space.*

Example 1.0.1. Let Ω be the set of outcomes of a fair coin toss. Then the function $\mu(H) = 0.5$ and $\mu(T) = 0.5$ defines a distribution. □

Example 1.0.2. For a number n, let $\pi(n) = 2^{-n}$. Then $(\mathbb{N} - \{0\}, \pi)$ forms a distribution. □

Let (Ω, F) be a discrete probability space for the subsequent discussion.

A *random variable* is a function $X : \Omega \to \Gamma$ where Γ is a discrete set. The probability distribution induced by X on Γ, denoted p_X, is defined for every $x \in \Gamma$ as

$$p_{_X}(x) = F\{\omega \in \Omega \mid X(\omega) = x\}.$$

This can be abbreviated as

$$p_{_X}(x) = F(X^{-1}(x)).$$

Thus X imposes a probability structure on the image Γ using the probability distribution on the domain Ω.

Example 1.0.3. Consider a game of darts, where the board has 3 bands, red, blue and green from outside in, and the bull's eye is marked in black. Let the score a player gets be 10, 20, 10 and 40, respectively for dart throws on red, blue, green and black. Assume that the probability that a dart falls on red is 0.5, blue is 0.2, green is 0.2 and bull's eye is 0.1.

This can be abstracted by the random variable: $X(\text{red}) = 10$, $X(\text{blue}) = 20$, $X(\text{green}) = 10$ and $X(\text{black}) = 40$.

Then after a throw, $p_X(10) = F(\{\text{ red, green }\}) = 0.7$. □

We now would like to define multidimensional probability distributions induced by random variables. Consider a distribution F_2 defined on $\Omega \times \Omega$.

Definition 1.0.4. Let $X : \Omega \to \Gamma$ and $Y : \Omega \to \Pi$ be two random variables. The *joint distribution* $p_{X,Y} : \Omega \times \Omega \to \Gamma \times \Pi$ is defined as

$$p_{X,Y}(\gamma, \pi) \quad = F_2\left((X^{-1}(\gamma), Y^{-1}(\pi))\right).$$

It is possible to define marginal distributions

$$p_{Y|X}(\pi \mid \gamma) = \frac{p_{X,Y}(\gamma, \pi)}{p_X(\gamma)}.$$

That is, $p_{Y|X}(\cdot \mid \gamma)$ is the probability distribution which is produced by scaling down the whole of the space Γ to a particular outcome γ. It follows that

$$p_{X,Y}(\gamma, \pi) \quad = p_X(\gamma)\, p_{Y|X}(\pi \mid \gamma) \quad = p_Y(\pi)\, p_{X|Y}(\gamma \mid \pi).$$

Example 1.0.5. Let $\Omega = \{H, T\}$ represent the outcome of a coin toss. We consider the sample space $\Omega \times \Omega$ of two coin tosses. Let F_2 be defined as $F_2(T, T) = 0.1$, $F_2(H, T) = 0.2$, $F_2(T, H) = 0.3$ and $F_2(H, H) = 0.4$.

Let X be the map $X(T) = 1$, $X(H) = 2$, and Y be the map $Y(H) = 10$, $Y(T) = 20$. The following table represents a joint distribution X, Y. (That is, the first coin toss is "scored" according to X and the second according to Y.)

	(0.25,0.75)	(0.33, 0.67)	
20	0.3	0.4	(0.42, 0.58)
10	0.1	0.2	(0.33, 0.67)
	1	2	

□

An important notion which characterizes the theory of probability, is the idea of *independence*. Independence tries to capture the idea that the occurrence of an event does not give any information about the occurrence of another. The mathematical way to capture this notion is as follows. Two outcomes are said to be independent if $F_2(\omega_1, \omega_2) = F(\omega_1)F(\omega_2)$. Similarly, two events A and B are independent if $F(A \cap B) = F(A)F(B)$. We extend this to the concept of two random variables being independent.

Definition 1.0.6. Two random variables $X : \Omega \to \Gamma$ and $Y : \Omega \to \Pi$ are *independent* if

$$p_{X,Y}(\gamma, \pi) = p_{X}(\gamma) p_{Y}(\pi).$$

Alternatively, we can say that X is independent of Y if $p_{X|Y} = p_{X}$. We can verify that if X is independent of Y, then Y is independent of X. We can think of this as the most basic instance of "symmetry of information".

Example 1.0.7. It is easily verified that the joint distribution in the previous example does not define lead to independent X and Y. Let Ω, X and Y be as before. Let F_2 be defined as $F_2(TT) = 0.01$, $F_2(TH) = 0.09$, $F_2(HT) = 0.09$, $F_2(HH) = 0.81$. This distribution is the product distribution generated by

$$F = \begin{pmatrix} T & H \\ 0.1 & 0.9 \end{pmatrix}.$$

We can verify that X and Y are independent of each other in this distribution.

	(0.1,0.9)	(0.1, 0.9)	
20	0.09	0.81	(0.1, 0.9)
10	0.01	0.09	(0.1, 0.9)
	1	2	

□

For more than two random variables, we can define various degrees of independence.

Definition 1.0.8. Random variables $X_1, X_2, \ldots, X_n$ are said to be *mutually independent* if

$$p_{X_1, X_2, \ldots, X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1), p_{X_2}(x_2) \ldots p_{X_n}(x_n).$$

Mutual independence is the strongest form of independence among n random variables. Another frequently useful notion is the weaker notion of *pairwise independence.*

Definition 1.0.9. Random variables $X_1, X_2, \ldots, X_n$ are said to be *pairwise independent* if every pair of distinct random variables among them are independent.

Mutual independence implies pairwise independence, but not conversely. An example to show that pairwise independence does not imply mutual independence is shown below, using the properties of the parity function.

Example 1.0.10. Consider two bits b_0 and b_1 produced by flips of a fair coin, and designating T with 1 and H with 0. Consider b_2 defined as the parity (XOR) of b_0 and b_1. It can be verified that any pair (b_i, b_j) of distinct bits are independent, but the three variables are not mutually independent. □

With this basic set of definitions from probability, we can now define the notion of *entropy* of a random variable.

Definition 1.0.11. Let (Ω, P) be a discrete probability space, and $X : \Omega \to \Gamma$ be a random variable. The *entropy* of the random variable X is defined to be

$$H(X) = -\sum_{x \in \Gamma} p_{_X}(x) \, \log p_{_X}(x),$$

where we adopt the convention that $0 \log 0 = 0$.

We can think of the entropy of a random variable as follows. If we assign an optimal coding scheme for the image $X(\Omega)$, where the higher the probability of a point, the fewer the bits we use in its encoding, we would use $\log \frac{1}{P(x)}$ bits to represent a point x. The entropy then is the weighted average of the length of this encoding scheme. Thus the entropy is the expected length of an optimal encoding of the random variable X, where X is distributed according to p_X.

Once we are familiar with the notion of probabilities induced by a random variable, we can drop the subscript from the probability. This is done to ease the burden of notation, when there is no confusion regarding which random variable we are talking about.

In the case where Γ consists of two symbols, we have the binary entropy function, $H(X) = -p \log p - (1-p) \log(1-p)$. We denote this as $h(p)$.

For two random variables X and Y, we can define the notions of joint entropy, and conditional entropies.

Definition 1.0.12. The *joint entropy* of two random variables X and Y is defined as

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) = -E \log p_{_{X,Y}}.$$

The *conditional entropy* of Y given X is defined by

$$H(Y \mid X) = -\sum_{x,y} p_{_{X,Y}}(x,y) \log p_{_{Y|X}}(y \mid x) = -E_{p_{_{X,Y}}} \log p_{_{Y|X}}.$$

Note the asymmetry in the last definition. We can understand this condition better by writing $p(x,y)$ as $p(x)p(y \mid x)$. Then

$$H(Y \mid X) = -\sum_{x,y} p(x) p(y \mid x) \log p(y \mid x).$$

This summation can now be separated into two sums, since x and y do not vary together.

$$H(Y \mid X) = \sum_{x} p(x) \left[-\sum_{y} p(y \mid x) \log p(y \mid x) \right].$$

The inner term is the entropy $H(Y \mid X = x)$. Thus

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x).$$

Definition 1.0.13. The information about X contained in Y is defined as $I(X;Y) = H(X) - H(X \mid Y)$.

Then, we have the property of *symmetry of information.*

Lemma 1.0.14. *$I(X;Y) = I(Y;X)$.*

Proof.

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= -\sum_x p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y) \\ &= \sum_x p(x) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \end{aligned}$$

We know that the probability of a point x, namely $p(x)$ is the sum of all $p(x,y)$ where y takes all values in Π. Thus, we can write the above sum as

$$\begin{aligned} \sum_x \sum_y p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} &= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \quad \text{[Notation]} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

By the symmetry of the resultant expression, $I(X;Y) = I(Y;X)$. □

An important property of the logarithm function is its concavity. The secant between two points on the graph of a *concave* function always lies below the graph. This characterizes concavity. A consequence of this fact is Jensen's Inequality. This is a very convenient tool in the analysis of convex or concave functions.

Theorem 1.0.15 (Jensen's Inequality)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable with $E[X] < \infty$, and $f : \mathbb{R} \to \mathbb{R}$ be a concave function. Then,*

$$f(E[X]) \geq E[f(X)].$$

Jensen's inequality can hence be used to perform the following kind of inequality:

$$\log \left[\sum p_i x_i \right] \geq \sum p_i \log x_i.$$

Usually, the left side is easier to estimate. From the geometric observation that the tangent to the graph of a concave function lies above the graph, we have the following useful upper bound on the logarithm function.

Theorem 1.0.16 (Fundamental Inequality)**.** *For any $a > 0$, we have*

$$\ln a \leq a - 1.$$

To illustrate an application of Jensen's inequality, we will prove that the entropy of an n-dimensional probability distribution is at most $\log n$.

Lemma 1.0.17. *Let $P = (p_0, p_2, \ldots, p_{n-1})$ be an n-dimensional probability distribution. Then $H(P) \leq \log n$.*

Proof.

$$\begin{aligned} H(P) &= \sum_{i=0}^{n-1} \left[p_i \log \frac{1}{p_i} \right] \\ &\leq \log \sum_{i=0}^{n-1} \frac{p_i}{p_i} && \text{[log is concave, and Jensen's inequality.]} \\ &= \log \sum_{i=0}^{n-1} 1 \\ &= \log n. \end{aligned}$$

□

Similarly, the fundamental inequality gives us a lower bound on the entropy, proving that it is non-negative.

Lemma 1.0.18. *Let $P = (p_0, p_1, \ldots, p_{n-1})$ be an n-dimensional probability distribution. Then $H(P) \geq 0$.*

Proof.

$$\begin{aligned} H(P) &= \sum_{i=0}^{n-1} p_i \log \frac{1}{p_i} \\ &\geq \sum_{i=0}^{n-1} p_i \left[1 - p_i\right] && \text{[Fundamental Inequality].} \end{aligned}$$

Both p_i and $1 - p_i$ are non-negative terms, since they are coordinates of a probability vector. $H(P)$ is lower bounded by a sum of nonnegative terms, thus, $H(P) \geq 0$. □

We can push the above analysis a bit further and get the result that $H(P) = 0$ if and only if P is a deterministic distribution. It is easily verified that if P is a deterministic distribution, then $H(P) = 0$. We now prove the converse.

Lemma 1.0.19. *Let $P = (p_0, p_1, \ldots, p_{n-1})$ be an n-dimensional probability distribution. Then $H(P) = 0$ only if P is a deterministic distribution.*

Proof. By the fundamental inequality, we have

$$H(P) \geq \sum_{i=0}^{n-1} p_i \left[1 - p_i\right] \geq 0.$$

Assume $H(P) = 0$. Then, the above sum of non-negative terms is 0, thus each of the constituent summands is equal to 0. This can happen only if $p_i = 1$ or $p_i = 0$ for each i. Observing that P is a probability distribution, thus exactly one p_i is 1, is enough to prove the result. □

Similarly, from symmetry of information and the fundamental inequality, we can prove that $I(X:Y) \geq 0$.

For, we have the following.

$$\begin{aligned}
I(X:Y) &= -\sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\
&\geq \sum_{x,y} p(x,y) \left[1 - \frac{p(x)p(y)}{p(x,y)}\right] && \text{(by Fundamental Inequality for } -\log.) \\
&= \sum_{x,y} p(x,y) - \sum_{x,y} p(x)p(y) \\
&= 1 - \sum_{x} p(x) \sum_{y} p(y) \\
&= 1 - 1 = 0.
\end{aligned}$$

1.1 Majorization*

We will just introduce a theory that is very useful in the study of entropy, how entropy changes when the probability distribution changes. Consider probability distributions on a set of n elements. We have proved that the entropy of the uniform distribution is maximal, and is equal to $\log n$. Similarly, we have proved that the entropy of a "deterministic" distribution, where one of the events has probability 1, and the rest have probability 0, is minimal, and equal to 0 (take $0 \log 0 = 0$ by convention.)

This says the following. The space of probability distributions on n elements can be seen as a convex set, with vertices being the deterministic distributions $(1, 0, 0, ..., 0)$, $(0, 1, 0, \dots, 0)$, $\dots$, $(0, 0, 0, \dots, 1)$. The entropy is maximal at the centroid of this set - this point corresponds to the uniform distribution. It then decreases outwards towards the vertices, and reaches 0 at each of the vertices.

How do we compare the entropy at two arbitrary points within this convex set? Can we determine some easily verifiable property of the probability distributions and use that to say qualitatively which distribution has greater entropy?

There is a powerful theory of majorization which can be used for this purpose. Majorization is a comparison criterion for two n-dimensional vectors. A vector $\vec{x}$ is said to be majorized by another vector $\vec{y}$ if, informally, $\vec{y}$ is more "equitably distributed" than $\vec{x}$.

Definition 1.1.1. Let $\vec{x}$ and $\vec{y}$ be two non-negative n-dimensional vectors. Let $\vec{a}$ be $\vec{x}$ sorted in descending order of coordinate values, and $\vec{b}$ be $\vec{y}$ sorted in descending order of coordinate values. Then, $\vec{x}$ is *majorized* by $\vec{y}$ and we write $\vec{x} \leq \vec{y}$ if the following hold.

$$\begin{aligned}
a_0 &\geq b_0 \\
a_0 + a_1 &\geq b_0 + b_1 \\
&\cdots \\
a_0 + a_1 + \cdots + a_n - 1 &= b_0 + b_1 + \cdots + b_{n-1}.
\end{aligned}$$

For example, $(1, 0, 0) \leq (1/2, 0, 1/2) \leq (1/3, 1/3, 1/3)$.

Thus, majorization provides an way to compare probability vectors. If a probability vector majorizes another probability vector, then their entropies can be compared. This is because an n-dimensional entropy function has a property called *Schur concavity.*

Definition 1.1.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *Schur concave* if $f(\vec{x}) \geq f(\vec{y})$ whenever $\vec{x} \leq \vec{y}$.

Shannon-entropy is Schur concave, as can be seen by an easy application of the following theorem.

Theorem 1.1.3 (Schur)**.** *A symmetric, continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is Schur-concave if and only if the projection of f onto each coordinate is a continuous concave function.*

The theorem connecting these concepts enables us to compare the entropies of probability vectors in an easy manner. A probability vector $\vec{x}$ has greater entropy than $\vec{y}$ if $\vec{x} \leq \vec{y}$.

Note that, majorization is just a sufficient condition for comparing entropies. There are vectors which do not majorize one another, with different entropies: For example, the 3-dimensional probability vector $(0.6, 0.25, 0.15)$ has greater entropy than $(0.5, 0.5, 0)$ even though they do not majorize one another.

1.2 Kullback-Leibler Divergence

Majorization was a qualitative notion of comparing two probability distributions. A very useful quantitative notion to compare two probability measures is known as the Kullback-Leibler divergence.

Definition 1.2.1. Let P and Q be two n-dimensional probability distributions. The *Kullback-Leibler Divergence* of Q from P is defined as

$$D(P||Q) = \sum_{i=0}^{n-1} P_i \log \frac{Q_i}{P_i}.$$

We can interpret this as follows. Note that

$$D(P||Q) = -\sum_{i=0}^{n-1} P_i \log P_i + \sum_{i=0}^{n-1} P_i \log Q_i = H(P) - \left(-\sum_{i=0}^{n-1} P_i \log Q_i \right).$$

This form is amenable to some interpretation. Suppose a sample space is distributed according to P. We mistakenly encode the space as though the space were distributed according to Q. Then, the divergence $D(P||Q)$ may be interpreted as the ‘coding inefficiency rate’ of encoding the space with respect to Q when the optimal rate would have been achieved with respect to P.

The KL-divergence does not have many properties of a Euclidean distance, and hence is not a satisfactory notion of ‘distance’ between probability vectors. However, it is remarkably useful.

Note: For example, the universal integrable test for a computable probability measure P in the previous chapter was the total KL divergence of M from P on any string x. Also, the deficiency test for the weak law of large numbers was approximately the KL divergence between the uniform distribution and p_x.

We conclude our discussion by proving some basic facts about the KL-divergence. We prove that $D(P||Q) = 0$ if and only if $P = Q$. One direction is easy: if $P = Q$, then $D(P||Q) = 0$. The converse is slightly tricky.

Lemma 1.2.2. *For finite positive probability distributions P and Q, $P = Q$ only if $D(P||Q) = 0$.*

Proof. First, we prove that $D(P||Q) \geq 0$. We will analyze the calculations needed in this result to prove the lemma.

First, we see that

$$\begin{aligned} D(P||Q) &= \sum p_i \log \frac{p_i}{q_i} \\ &= \sum p_i \left[- \log \frac{q_i}{p_i} \right] \\ &\geq \sum p_i \left[1 - \frac{q_i}{p_i} \right] && \text{[Fundamental Inequality]} \\ &= \sum p_i - \sum q_i \\ &= 0. && [P \text{ and } Q \text{ are probabilities.}] \end{aligned}$$

Then $D(P||Q) = 0$ only if the inequality between lines 2 and 3 is an equality. We know that

$$- \log \frac{q_i}{p_i} - \left[1 - \frac{q_i}{p_i} \right] \geq 0,$$

hence every summand

$$p_i \left[- \log \frac{q_i}{p_i} \right]$$

in line 2 is at least the corresponding summand

$$p_i \left[1 - \frac{q_i}{p_i} \right]$$

.

Thus line 2 and line 3 are equal only if the corresponding summands are equal. Thus, for every i, $\frac{q_i}{p_i} = 1$, proving that $P = Q$. □