ECE 7680
Lecture 2 – Definitions and Basic Facts

Objective: To learn a bunch of definitions about entropy and information measures that will be useful through the quarter, and to present some simple but important theorems: Jensen's inequality, and the information inequality

In this lecture a bunch of definitions and a few simple theorems are presented. While these might be somewhat bewildering at first, you should just hang on and dig in. There is a sort of "algebra of informational quantities" that must be presented as preliminary material before we can get much further.

The binary entropy function

We saw last time that the entropy of a random variable X is

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

Suppose X is a binary random variable,

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Then the entropy of X is

$$H(X) = -p \log p - (1-p) \log(1-p)$$

Since this depends on p, this is also written sometimes as $H(p)$. Plot. Observe: **concave** function of p. (What does this mean?) $H(0) = 0$, $H(1) = 0$. Why? Where is the max?

More generally, the entropy of a binary discrete random variable with probability p is written as either $H(X)$ or $H(p)$.

Joint entropy

Often we are interested in the entropy of pairs of random variables (X, Y). Another way of thinking of this is as a vector of random variables.

Definition 1 If X and Y are jointly distributed according to $p(X, Y)$, then the **joint entropy** $H(X, Y)$ is

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

or

$$H(X, Y) = -E \log p(X, Y)$$

□

Definition 2 If $(X, Y) \sim p(x, y)$, then the **conditional entropy** $H(Y|X)$ is

$$H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) = -E_{p(x,y)} \log p(y|x)$$

This can also be written in the following equivalent (and also useful) ways:

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

□

Theorem 1 *(*chain rule*)*

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

Interpretation: The uncertainty (entropy) about both X and Y is equal to the uncertainty (entropy) we have about X, plus whatever we have about Y, given that we know X.

Proof This proof is very typical of the proofs in this class: it consists of a long string of equalities. (Later proofs will consist of long strings of inequalities. Some people make their livelihood out of inequalities!)

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

This can also be done in the following streamlined manner: Write

$$\log p(X,Y) = \log p(X) + \log p(Y|X)$$

and take the expectation of both sides. □

We can also have a joint entropy with a conditioning on it, as shown in the following corollary:

Corollary 1

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

The proof is similar to the one above. (This is a good one to work on your own.)

Relative entropy and mutual information

Suppose there is a r.v. with true distribution p. Then (as we will see) we could represent that r.v. with a code that has average length $H(p)$. However, due to incomplete information we do not know p; instead we assume that the distribution of the r.v. is q. Then (as we will see) the code would need more bits to represent the r.v. The difference in the number of bits is denoted as $D(p\|q)$. The quantity $D(p\|q)$ comes up often enough that it has a name: it is known as the **relative entropy**.

Definition 3 The **relative entropy** or **Kullback-Leibler distance** between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = E_p \log\frac{p(X)}{q(X)}$$

□

Note that this is not symmetric, and the q (the second argument) appears only in the denominator.

Another important concept is that of *mutual information*. How much information does one random variable tell about another one. In fact, this perhaps the central idea in much of information theory. When we look at the output of a channel, we see the outcomes of a r.v. What we want to know is what went into the channel — we want to know what was sent, and the only thing we have is what came out. The channel coding theorem (which is one of the high points we are trying to reach in the class) is basically a statement about mutual information.

Definition 4 Let X and Y be r.v.s with joint distribution $p(X,Y)$ and marginal distributions $p(x)$ and $p(y)$. The **mutual information** $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution:

$$\begin{aligned} I(X;Y) &= D(p(x,y)\|p(x)p(y)) \\ &= \sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

□

Note that when X and Y are independent, $p(x,y) = p(x)p(y)$ (definition of independence), so $I(X;Y) = 0$. This makes sense: if they are independent random variables then Y can tell us nothing about X.

An important interpretation of mutual information comes from the following.

Theorem 2 $I(X;Y) = H(X) - H(X|Y)$

Interpretation: The information that Y tells us about X is the reduction in uncertainty about X due to the knowledge of Y.

Proof

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

□

Observe that by symmetry

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

That is, Y tells as much about X as X tells about Y. Using $H(X,Y) = H(X) + H(Y|X)$ we get

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

The information that X tells about Y is the uncertainty in X plus the uncertainty about Y minus the uncertainty in both X and Y. We can summarize a bunch of statements about entropy as follows:

$$\boxed{\begin{array}{l} I(X;Y) = H(X) - H(X|Y) \\ I(X;Y) = H(Y) - H(Y|X) \\ I(X;Y) = H(X) + H(Y) - H(X,Y) \\ I(X;Y) = I(Y;X) \\ I(X;X) = H(X) \end{array}}$$

More than two variables

It will be important to deal with more than one or two variables, and a variety of "chain rules" have been developed for this purpose. In each of these, the sequence of r.v.s $X_1, X_2, \ldots, X_n$ are drawn according to the joint distribution $p(x_1, x_2, \ldots, x_n)$.

Theorem 3 *The joint entropy of $X_1, X_2, \ldots, X_n$ is*

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

Proof Observe that

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2|X_1) \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \qquad (*) \\ &\vdots \\ H(X_1, X_2, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \end{aligned}$$

An alternate proof can be obtained by observing that

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)$$

and taking an expectation. □

We sometimes have two variables that we wish to consider, both conditioned upon another variable.

Definition 5 The **conditional mutual information** of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= E \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

□

In other words, it is the same as mutual information, but everything is conditioned upon Z.

The chain rule for entropy leads us to a *chain rule for mutual information.*

Theorem 4

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1).$$

Proof

$$\begin{aligned} I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1, Y) \\ &= \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1). \end{aligned}$$

□

(Skip conditional relative entropy for now.)

Convexity and Jensen's inequality

A large part of information theory consists in finding *bounds* on certain performance measures. The analytical idea behind a bound is to substitute a complicated expression for something simpler but not exactly equal, known to be either greater or smaller than the thing it replaces. This gives rise to simpler statements (and hence gain some insight), but usually at the expense of precision. Knowing when to use a bound to get a useful results generally requires a fair amount of mathematical maturity and experience.

One of the more important inequalities we will use throughout information theory is Jensen's inequality. Before introducing it, you need to know about convex and concave functions.

Definition 6 A function $f(x)$ is said to be **convex** over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2)) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

A function is **strictly convex** if equality holds only if $\lambda = 0$ or $\lambda = 1$. □

To understand the definition, recall that $\lambda x_1 + (1-\lambda)x_2)$ is simply a line segment connecting x_1 and x_2 (in the x direction) and $\lambda f(x_1) + (1-\lambda)f(x_2)$ is a line segment connecting $f(x_1)$ and $f(x_2)$. Pictorially, the function is convex if the *function lies below the straight line segment connecting two points*, for any two points in the interval.

Definition 7 A function f is **concave** if $-f$ is convex. □

You will need to keep reminding me of which is which, since when I learned this, the nomenclature was "convex ∪" and "convex ∩".

Example 1

Convex x^2, e^x, $|x|$, $x \log x$.

Concave: $\log x$, $\sqrt{x}$

□

One reason why we are interested in convex functions is that it is known that *over the interval of convexity there is only one minimum.* This can strengthen many of the results we might want.

Theorem 5 *If f has a second derivative which is non-negative (positive) everywhere, then f is convex (strictly convex).*

Proof The Taylor-series expansion of f about the point x_0 is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x^*)(x - x_0)^2$$

If $f''(x) \geq 0$ the last term is non-negative.

Let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and let $x = x_1$. Then

$$f(x_1) \geq f(x_0) + f'(x_0)[(1 - \lambda)(x_1 - x_2)].$$

Now let $x = x_2$ and get

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]$$

Multiply the first by λ and the second by $1-\lambda$, and add together to get the convexity result. □

We now introduce **Jensen's inequality**.

Theorem 6 *If f is a convex function and X is a r.v. then*

$$Ef(X) \geq f(EX).$$

Put another way,

$$\sum_x p(x)f(x) \geq f\left(\sum_x p(x)x\right)$$

If f is strictly convex then equality in the theorem implies that $X = EX$ w.p. 1.

If f is concave then

$$Ef(X) \leq f(EX).$$

The theorem allows us (more or less) to pull a function outside of a summation in some circumstances.

Proof The proof is by induction. When X takes on two values the inequality is

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2).$$

This is true by the definition of convex functions.

Inductive hypothesis: suppose the theorem is true for distributions with $k - 1$ values. Then let $p'_i = p_i/(1 - p_k)$ for $i = 1, 2, \ldots, k - 1$,

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\
&\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) \\
&\geq f\left(p_x x_k + (1 - p_k) \sum_{i=1}^{k-1} p'_i x_i\right) \\
&= f\left(\sum_{i=1}^{k} p_i x_i\right)
\end{aligned}$$

□

There is another inequality that got considerable use (in many of the same ways as Jensen's inequality) way back in the dark ages when I took information theory. I may refer to it simply as the **information inequality**.

Theorem 7 $\log x \leq x - 1$, *with equality if and only if* $x = 1$.

This can also be generalized by taking the line at different points along the function.

With these simple inequalities we can now prove some facts about some of the information measures we defined so far.

Theorem 8 $D(p\|q) \geq 0$, *with equality if and only if* $p(x) = q(x)$ *for all* x.

Proof

$$\begin{aligned} -D(p\|q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} & \\ &= \sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_x p(x) \frac{q(x)}{p(x)} \qquad \text{(Jensen}'\text{s)} \\ &= \log \sum_x q(x) = 0 \end{aligned}$$

Since log is a strictly concave function, we have equality if and only if $q(x)/p(x) = 1$.

□

Proof Here is another proof using the information inequality:

$$\begin{aligned} -D(p\|q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\leq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \qquad (\log x \leq x - 1) \\ &= \sum_x q(x) - p(x) = 0. \end{aligned}$$

□

Corollary 2 *Mutual information is positive:*

$$I(X;Y) \geq 0,$$

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

Let $\mathcal{X}$ be the set of values that the random variable X takes on and let $|\mathcal{X}|$ denote the number of elements in the set. For discrete random variables, *the uniform distribution over the range* $\mathcal{X}$ *has the maximum entropy.*

Theorem 9 $H(X) \leq \log |\mathcal{X}|$, *with equality iff* X *has a uniform distribution.*

Proof Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform distribution and let $p(X)$ be the distribution for X. Then

$$D(p\|u) = \sum_x p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X).$$

□

Note how easily this optimizing value drops in our lap by means of an inequality. There is an important principle of engineering design here: if you can show that some performance criterion is upper-bounded by some function, then show how to achieve that upper bound, you have got an optimum design. No calculus required!

The more we know, the less uncertainty there is:

Theorem 10 *Condition reduces entropy:*

$$H(X|Y) \leq H(X),$$

with equality iff X and Y are independent.

Proof $0 \leq I(X;Y) = H(X) - H(X|Y)$. □

Theorem 11

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

with equality if and only if the X_i are independent.

Proof By the chain rule for entropy,

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

□