COMP 761: Quantum Information Theory **Winter 2009**

Lecture 2 — January 8

Lecturer: Patrick Hayden *Scribe: Omar Fawzi*

In this lecture, we introduce some functions that will be very useful to tackle information theory questions. To fix the notations, upper case letters will denote random variables (e.g. X, Y) and their lower case will correspond to a possible value of this random variable (e.g x, y). Moreover, the random variables considered will take values in finite sets usually called χ. The distribution of a random variable X will be denoted by $\mathbb{P}(X = x) = p(x)$.

Recall from last lecture that we defined the entropy of a random variable X as

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

2.1 Conditional entropy and mutual information

The entropy is meant to measure the uncertainty in a realization of X. Now, we want to quantify how much uncertainty does the realization of a random variable X have if the outcome of another random variable Y is known.

We define the *conditional entropy* as

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

Here $p(y)$ is the marginal distribution of Y, so $p(y) = \sum_x p(x,y)$. It is easy to see that the conditional entropy can be written as $H(X|Y) = H(X,Y) - H(Y)$. For example, if $X = Y$, then $H(X|Y) = 0$.

Observe that because $H(X|Y = y) \geq 0$, the conditional entropy is always positive: $H(X|Y) \geq 0$. In other words

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

which seems intuitive, as the uncertainty of the whole has to be at least as big as the uncertainty of a part. But this remark is important as this will not be true for quantum information. In fact, in quantum information, $H(X|Y) < 0$ is possible.

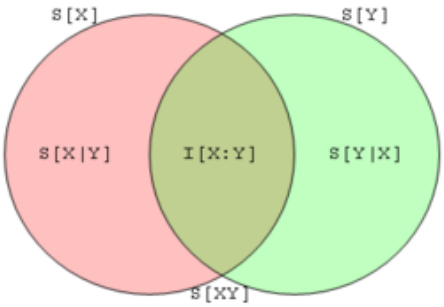

Figure 2.1. Representation of the entropy functions family (taken from Wikipedia entry on conditional entropy)

Now we introduce a complementary quantity, that measures the common uncertainty between X and Y, the *mutual information* is defined as

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

One can see from the last expression that the mutual information is symmetric in X and Y. This symmetry means that this notion of uncertainty has the property that the information we gain about X when knowing Y is the same as the information we gain about Y knowing X. To illustrate this property, consider X uniformly distributed on the set $\{1, \ldots, 2^k\}$, and Y the parity of X, then knowing X determines Y, so it gives exactly 1 bit of information. Moreover, knowing Y gives the parity of X, so it also gives 1 bit of information about X.

The Venn diagram in figure 2.1 illustrates the relations between the different functions we introduced.

2.2 Relative entropy

Another function that will be useful is the *relative entropy*, which is a measure of closeness between probability distributions over the same set

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Note that the relative entropy is not symmetric in p and q. We give some examples of relative entropies.

If $u(x)$ is the uniform distribution on χ, then

$$D(p||u) = \sum_x p(x) \log \frac{p(x)}{1/|\chi|} = \log |\chi| - H(X)$$

where X has distribution $p(x)$.

Moreover, one can write the mutual information as a relative entropy.

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

Theorem 2.2.1. *$D(p||q) \geq 0$ with equality iff $p = q$.*

Proof:

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= -\frac{1}{\ln 2} \sum_x p(x) \ln \frac{q(x)}{p(x)} \\ &\geq -\frac{1}{\ln 2} \sum_x p(x) (\frac{q(x)}{p(x)} - 1) \end{aligned}$$

where we used the inequality $\ln x \leq x - 1$, which is a consequence of the concavity of ln. Now using the fact that $p(x)$ and $q(x)$ sum to one, we get

$$D(p||q) \geq 0.$$

Equality is achieved iff for all x, $\ln \frac{q(x)}{p(x)} = \frac{q(x)}{p(x)} - 1$, which is equivalent to $p(x) = q(x)$ for all x. □

Some consequences:

- As $D(p||u) = \log|\chi| - H(X)$, where u is the uniform distribution on χ, then for any X taking values in χ, $H(X) \leq \log|\chi|$. Moreover, by the equality condition, the uniform distribution is the unique maximum of the entropy function among the distributions on χ.
- As we expressed the mutual information as a relative entropy, $I(X:Y) = H(X) + H(Y) - H(X,Y) \geq 0$, which is sometimes called the *subaddtivity* property. Furthermore, equality happens if and only if $p(x,y) = p(x)p(y)$ which means that X and Y are independent. This means that $I(X:Y)$ captures all kinds of relations between X and Y, not only linear dependencies as the the correlation coefficient for example.
- By rewriting the positivity of mutual information as $H(X|Y) \leq H(X)$, one can interpret it as "conditioning reduces entropy"

2.3 Some useful equalities

When given many random variable, it is often useful to decompose the entropy of the joint distribution in the following way using the *chain rule*:

$$H(X_1, \ldots, X_n|Y) = \sum_{j=1}^{n} H(X_j|Y, X_1, \ldots X_{j-1}).$$

One can also give a similar formula for the mutual information, but first we have to define the *conditional mutual information*:

$$\begin{aligned} I(X:Y|Z) &= \sum_z p(z) I(X:Y|Z=z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \end{aligned}$$

Note that $I(X:Y|Z) \geq 0$.

Using this definition, we can decompose mutual information

$$I(X_1, \ldots, X_n : Y) = \sum_{j=1}^{n} I(X_j : Y|X_1, \ldots X_{j-1}).$$

2.4 Data processing inequality

Suppose $X, Y, Z \sim p(x, y, z)$. We can write

$$\begin{aligned} p(x, y, z) &= p(x, y)p(z|x, y) \\ &= p(x)p(y|x)p(z|x, y) \end{aligned}$$

This gives a way to sample (x_0, y_0, z_0) from the joint distribution $p(x, y, z)$, by first taking x_0 with distribution $p(x)$, then y_0 with distribution $p(y|x_0)$, then z_0 with distribution $p(z|x_0, y_0)$.

We will consider a particular kind of relation between X, Y, Z in which Z depends on X only through Y.

Definition 2.4.1. *X, Y, Z is a Markov chain if*

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

We write $X - Y - Z$.

For example, X, Y, Z is a Markov chain if Z is a deterministic function of Y.

Note that X, Y, Z is a Markov chain if and only if X and Z are conditionally independent given Y, or in other words, $I(X : Z|Y) = 0$. In fact, if $X - Y - Z$, then

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x, y)}{p(y)}p(z|y) = p(x|y)p(z|y).$$

Defined in this way, we see that there is no directionality in $X - Y - Z$.

Theorem 2.4.2 (Data processing inequality). *If $X - Y - Z$, then $I(X : Y) \geq I(X : Z)$. In words, post-processing Y cannot increase mutual information with X.*

Proof: Consider $I(X : Y, Z)$. On the one hand,

$$I(X : Y, Z) = I(X : Y) + I(X : Z|Y) = I(X : Y).$$

On the other hand,

$$I(X : Y, Z) = I(X : Z) + I(X : Y|Z) \geq I(X : Z)$$

as $I(X : Y|Z) \geq 0$. As a result, $I(X : Y) \geq I(X : Z)$. □