Quantum Information 101 for physicists

Class exercise 9

Joint and conditional distributions

Assume we have a channel XY that emits pairs of letters (x, y) with probability distribution $\{p(x,y)\}_{x\in X, y\in Y}$. Then we can discuss $H(XY) = -\sum_{xy} p(x,y)\log p(x,y)$, but also

$$\begin{aligned} H(X) &= -\sum_x p(x)\log p(x) = -\sum_x\sum_y p(x,y)\log\sum_{y'} p(x,y'), \\ H(Y|X) &= H(XY) - H(X) = \langle H(p(y\,|\,x))\rangle\,, \end{aligned}$$

and their equivalent $H(Y), H(X|Y)$. $H(X|Y), H(Y|X)$ are called **conditional entropies**. We can use them to define the mutual information,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(XY) = \left\langle \log\frac{p(x,y)}{p(x)p(y)}\right\rangle. \end{aligned}$$

The mutual information tells us how correlated X and Y are, that is, how much new information we can learn on X given an access to Y.

Example 1

We look at a specific example:

$$p(0,0) = \frac{1}{2}, p(0,1) = \frac{1}{4}, p(1,0) = \frac{1}{8}, p(1,1) = \frac{1}{8}.$$

For this distribution we compute the marginal probabilities:

$$p(x=0) = \frac{3}{4}, p(x=1) = \frac{1}{4}, p(y=0) = \frac{5}{8}, p(y=1) = \frac{3}{8},$$

and the conditional probabilities:

$$p(y|x=0) = \begin{cases} \frac{2}{3} & y=0 \\ \frac{1}{3} & y=1 \end{cases}, p(y|x=1) = \begin{cases} \frac{1}{2} & y=0 \\ \frac{1}{2} & y=1 \end{cases}.$$

We now compute

$$\begin{aligned} H(Y|X) &= H(XY) - H(X) \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{2}{8}\log\frac{1}{8} + \frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4} \\ &= -\frac{1}{4}\log 2 + \frac{3}{4}\log 3 \end{aligned}$$

We can now check explicitly that this gives us the average entropy over the conditional probabilities:

$$\begin{aligned} \langle \log p(y\,|\,x)\rangle &= p(x=0)H(Y|x=0) + p(x=1)H(Y|x=1) \\ &= \frac{3}{4}\left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right) + \frac{1}{4}\log 2 \\ &= -\frac{1}{4}\log 2 + \frac{3}{4}\log 3. \end{aligned}$$

We now prove this for the general case:

$$\begin{aligned} H(Y|X) &= H(XY) - H(X) \\ &= -\sum_{xy} p(x,y)\log p(x,y) + \sum_x p(x)\log p(x) \\ &= -\sum_{xy} p(x,y)\log p(x,y) + \sum_x \left(\sum_y p(x,y)\right)\log\left(\sum_{y'} p(x,y')\right) \\ &= -\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{\sum_{y'} p(x,y')} \\ &= -\sum_{xy} p(x,y)\log\frac{p(x,y)}{p(x)} = -\sum_x p(x,y)\log p(y|x) = \langle p(y\,|\,x)\rangle\,. \end{aligned}$$

at home you will show the same for the mutual information.

Example 2

Show that the entropy is **subadditive**: Assume we have a source XY with the probability distribution $\{p(x,y)\}_{x\in X, y\in Y.}$ Then show that

$$H(XY) \le H(X) + H(Y).$$

When does the inequality become an equality?

Solution

$$H(XY) = H(X) + H(Y|X) \le H(X) + H(Y),$$

which is an equality only when

$$\begin{aligned} H(Y|X) = H(XY) - H(X) &= H(Y) \\ \Rightarrow H(XY) &= H(X) + H(Y). \end{aligned}$$

The inequality becomes an equality when knowing X tells us nothing about Y,that is, when Y, X are independent. The subadditivity tells us that by having access to both X and Y, we can only benefit by reducing the minimal required codeword length, rather by decoding each of them separately. It is also extremely useful for proving results.

We say that two sequences $\vec{x}, \vec{y}$of length n are jointly $\delta-$typical if

$$\begin{aligned} 2^{-n(H(X)+\delta)} &\le p(\vec{x}) \le 2^{-n(H(X)-\delta)}, \\ 2^{-n(H(Y)+\delta)} &\le p(\vec{y}) \le 2^{-n(H(Y)-\delta)}, \\ 2^{-n(H(XY)+\delta)} &\le p(\vec{x},\vec{y}) \le 2^{-n(H(XY)-\delta)}, \end{aligned}$$

so both $\vec{x}, \vec{y}$ and the combination of $\vec{x}$ and $\vec{y}$ are probable enough with respect to δ.

The noisy channel coding theorem

In class we discussed communication over a noisy channel and channel capacity. We quickly review what we discussed: we study a channel on which Alice transfers information from a source X and on Bob receives Y. The channel is noisy, so the characters Alice inputs don't necessarily get to Bob, and the channel is defined by some conditional probability $p(y|x)$. Alice and Bob may agree on some code that will allow Bob to decode Alice's information correctly despite the noise, in which a codeword of length k is encoded as a word of length n. Such a code requires some redundancy, that is, $n > k$, and we define the rate of the code, R, as

$$R = \frac{k}{n}.$$

We now turn to bound R when k, n are very large. This will allow us to quantify the bound imposed on R in the case when Alice and Bob are willing to work as hard as possible to define the code and communicate it between them, so all of the restrictions come from the channel itself.

Say we expect $h(n)$ errors when the message is transferred on the channel. Then we require that the *Hamming distance* between any two codewords to be larger than or equal to $h(n)$, where the Hamming distance is the number of required changes to turn one word into the other. This way, even after $h(n)$ errors Bob would still be able to recognize the original codeword. This gives us a bound on the relation between number of codewords $2^k$and number of possible words 2^n. In class we studied the binary symmetric channel:

$$p(0|0) = p(1|1) = 1 - p, p(0|1) = p(1|0) = p.$$

For this channel, $h(n) = nH(p)$, and then we get

$$2^k 2^{nH(p)} \leq 2^n \Rightarrow R \leq 1 - H(p) \equiv C(p),$$

where $C(p)$ is the channel capacity.

This was an easy case to analyze, since the probabilities $p(y|x)$ were distributed the same for $x = 0, 1$. In the general case, Alice has some input source X and the channel is defined by the probability $p(y|x)$. Alice chooses n codewords by campling X^n, and tells them to Bob beforehand. She then transmits some codeword x^n and Bob receives y^n. Bob now checks if there is a codeword in X^n that is jointly typical with y^n. If he finds one, he decodes this word, and otherwise he decodes arbitrarily. In class you saw that in this strategy, the required code rate R turns to be

$$R = I(X;Y) - o(1).$$

This gives us a better understanding of $I(X;Y)$: it symbolizes the length of the code required for X given that Y is known, that is, the amount of information in X that we cannot learn from Y.

The channel capacity is then defined as

$$C \equiv \max_X I(X;Y).$$

Example 3

We now compute the capacity of an erasure channel:

$$p(0|0) = \frac{3}{4}, p(1|0) = 0, p(2|0) = \frac{1}{4},$$
$$p(0|1) = 0, p(1|0) = \frac{1}{2}, p(2|0) = \frac{1}{2}.$$

We define the source X,

$$p(x = 0) = 1 - p_x, p(x = 1) = p_x.$$

We calculate $p(y)$:

$$p(y = 0) = \frac{3}{4}(1 - p_x), p(y = 1) = \frac{1}{2}p_x, p(y = 2) = (1 - p_x)\frac{1}{4} + p_x\frac{1}{2} = \frac{1}{4}(1 + p_x),$$

so the mutual information

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= -\frac{3}{4}(1 - p_x)\log\left(\frac{3}{4}(1 - p_x)\right) - \frac{1}{2}p_x\log\left(\frac{1}{2}p_x\right) - \frac{1}{4}(1 + p_x)\log\left(\frac{1}{4}(1 + p_x)\right) \\ &+ (1 - p_x)H\left(\frac{1}{4}\right) + p_x\log 2 \end{aligned}$$

We may now find p_x that maximaizes the mutual information:

$$\frac{dI(X;Y)}{dp_x} = \frac{3}{4}p_x\left(\log\left(\frac{3}{4}(1 - p_x)\right) + 1\right) - \frac{1}{2}\left(\log\left(\frac{1}{2}p_x\right) + 1\right) - \frac{1}{4}\left(\log\left(\frac{1}{4}(1 + p_x)\right) + 1\right) + -H\left(\frac{1}{4}\right) + \log 2 = 0$$

This equation can be solved numerically to find the channel capacity.