Quantum Information 101 for physicists

Class exercise 8

We start with the core of information theory - Shannon's entropy. It deals with information in texts, random information sources, sending information between parties through noisy channels, and error correction, and is extremely usefull in understanding the behavior and abilities of computers when the amount of data becomes very very large. Later when we generalize it to quantum physics, we will get two things: First, like in classical computers, quantum information helps us study the behavior of quantum computers when there is a lot of data. Second, like you see in statistical mechanics, the running of information in the systems describes the physical behavior of the system. QI allows us (among other things) to measure entanglement, which is proven to be extremely important in quantum field theories and quantum many-body physics, by signaling phase transitions and interesting behaviors. But all of this is for the rest of the course.

Let us start from the classical case. We recall some definitions from class: Assume we have some source X that emits the letter x with probability $p(x)$, then the entropy of the source is

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

As in thermodynamics, the entropy measures the 'disorder' of the source, or the number of typical strings in the source. Let's see an example with three possible letters, $0, 1, 2$. We denote a channel by its probabilities in the form $(p(0), p(1), p(2))$:

$$\begin{aligned}
H\left((1,0,0)\right) &= -1\cdot\log(1) = 0,\\
H\left(\left(\frac{1}{2},\frac{1}{2},0\right)\right) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2,\\
H\left(\left(\frac{1}{2},\frac{1}{2}-\epsilon,\epsilon\right)\right) &= -\frac{1}{2}\log\frac{1}{2} - (\frac{1}{2}-\epsilon)\log(\frac{1}{2}-\epsilon) - \epsilon\log\epsilon \approx \log 2 - \epsilon\log\epsilon,\\
H\left(\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)\right) &= -3\cdot\frac{1}{3}\log\frac{1}{3} = \log 3.
\end{aligned}$$

We see that the most unexpected source, $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$, has an entropy of $\log 3$. We turn this into the first exercise of the day:

Exercise 1

Show that the maximum value of the entropy for a source of d letters is $\log d$, and that this value is obtained for the uniform probability.

Solution

We use Jensen's inequality:

$$\begin{aligned}
H(X) &= -\sum_x p(x)\log p(x)\\
&= \sum_x p(x)\log\frac{1}{p(x)}\\
&\leq \log\sum_x p(x)\frac{1}{p(x)}\\
&= \log d.
\end{aligned}$$

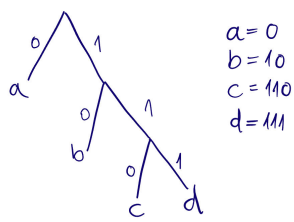
The inequality is true since the log function is concave – a weighted sum of logs cannot be larger than the log of the weighted sum. The inequality becomes equality when $p(x) = 1/d$ for all x, and then the entropy is maximal.

Prefix code

In class we saw an example to a prefix code, where we encode each letter according to its rate in the source, and saw that the length of the code was directly related to the entropy. We will now go over it again, but also see some other cases that don't fit this code perfectly and analyze them. We start by repeating the distribution from class:

$$p_a = \frac{1}{2}, p_b = \frac{1}{4}, p_c = \frac{1}{8}, p_d = \frac{1}{8}.$$

And the code:



The average length of a single character in this code would be

$$\lim_{n\to\infty} \frac{L}{n} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3,$$

where L is the length of the encoded string.

We now look at the entropy of the source:

$$\begin{aligned} H(X) &= \frac{1}{2} \cdot \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{8} \log 8 \\ &= \log 2 \left[\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 \right]. \end{aligned}$$

We see that the entropy here equals exactly the average coded character length of the code multiplied by the log of the basis length of a single physical character (bit, in our case).

We look at another simple example:

$$p_a = p_b = p_c = p_d = \frac{1}{4}.$$

Now, the prefix code is quite boring. We can just naively encode each letter by 2 bits, and then the average encoded character length would be 2. The entropy of the source is $H(X) = \log 4 = 2 \log 2$, which is again the average character length multiplied by $\log 2$. This leads us to interpret $H(X)$ as the average character length in the optimal encoding of the source.

We look at a less-structured source:

$$p_a = \frac{1}{2}, p_b = p_c = p_d = \frac{1}{6}.$$

The prefix code average length would be

$$\lim_{n\to\infty} \frac{L}{n} = \frac{1}{2} \cdot 1 + \frac{1}{6} \cdot (2 + 3 + 3) = 1\frac{5}{6} \approx 1.83,$$

and the entropy

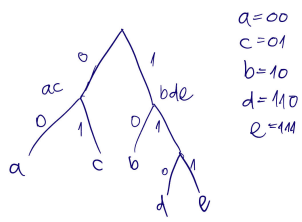
$$H(X) = \frac{1}{2} \log 2 + 3 \cdot \frac{1}{6} \log 6 \approx 1.79 \log 2.$$

Obviously, the prefix code is not the ideal method for this code, but we see that it is quite close. At home you will get closer to $\lim_{n\to\infty} \frac{L}{n} = 1.79$.

Finally, we look at the last code for today:

$$p_a = p_b = \frac{1}{3}, p_c = p_d = p_e = \frac{1}{9}.$$

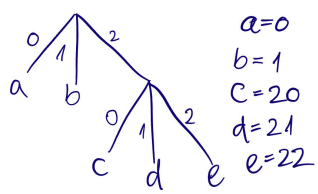
Now we have 5 letters, so naively we would need 3 bits to encode each of them. The prefix code would be:



with the average character kength and entropy:

$$\lim_{n\to\infty}\frac{L}{n} = \frac{1}{3}\cdot 2 + \frac{1}{3}\cdot 2 + \frac{1}{9}\cdot 2 + \frac{1}{9}\cdot 2\cdot 3 \approx 2.22,$$
$$H(X) = 2\cdot\frac{1}{3}\log 3 + 3\cdot\frac{1}{9}\log 9 \approx 2.11\log 2.$$

We didn't get $\lim_{n\to\infty}\frac{L}{n} = H(X)/\log 2$, which makes sense, since the code is not optimal - for example, we are decoding a and c in the same length. But let's look at a different situation: Assume that instead of bits, we would have 3-dimensional data units, which can be 0,1 or 2. In this case, we would have the following prefix code:



Now we see that this is more suitable for the code:

$$\lim_{n\to\infty}\frac{L}{n} = \frac{1}{3}\cdot 1\cdot 2 + \frac{1}{9}\cdot 2\cdot 3 = \frac{4}{3},$$
$$H(X) = 2\cdot\frac{1}{3}\log 3 + 3\cdot\frac{1}{9}\log 9 = \frac{4}{3}\log 3.$$

The above teeaches us that the entropy is not basis-dependent. It reflects how much information is in a text and how much we can compress it, but on a higher level, that doesn't regard the actual protocol or hardware we use for compressing.

Joint and conditional distributions

Assume we have a source 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

$$H(X) = -\sum_x p(x)\log p(x) = -\sum_x\sum_y p(x, y)\log\sum_{y'} p(x, y'),$$
$$H(X|Y) = H(XY) - H(Y),$$

and their equivalent $H(Y), H(Y|X)$. $H(X|Y), H(Y|X)$ are called **conditional entropies**, since they are related to the conditional probabilities $p(x|y), p(y|x)$:

$$\begin{aligned}
H(X|Y) &= -\sum_{xy} p(x, y)\log p(x, y) + \sum_y p(y)\log p(y)\\
&= -\sum_{xy} p(x, y)\log p(x, y) + \sum_y\sum_x p(x, y)\log p(y)\\
&= -\sum_{xy} p(x, y)\log\frac{p(x, y)}{p(y)} = -\sum_{xy} p(x, y)\log p(x|y) \geq 0,
\end{aligned}$$

Where the last inequality is true since all terms in the sum are greater or equal to 0. It is the average of $H(X|Y = y)$ for all ys. So the conditional entropy $H(X|Y)$ quantifies the information obtained from X given that Y is already known.

We can use them to define the mutual information,

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

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. Let's see some examples again, when X and Y both have two letters, 0 and 1:

- X, Y independent - $p_x(0) = p, p_y(0) = q$:

$$\begin{aligned} H(X) &= H(X|Y) = -p\log p - (1-p)\log(1-p), \\ H(Y) &= H(Y|X) = -q\log q - (1-q)\log(1-q), \\ H(XY) &= -pq\log pq - p(1-q)\log p(1-q) - (1-p)q\log(1-p)q - (1-p)(1-q)\log(1-p)(1-q) \\ &= H(X) + H(Y). \end{aligned}$$

  Then

$$I(X;Y) = H(X) - H(X|Y) = 0,$$

  and indeed, knowing Y will not tell us anything about X.

- X, Y completely dependent, for example $p(0,0) = p, p(1,1) = 1-p, p(1,0) = p(0,1) = 0$.

$$\begin{aligned} H(XY) = H(X) = H(Y) &= -p\log p - (1-p)\log p, \\ H(X|Y) = H(Y|X) &= 0, \\ I(X;Y) &= H(XY). \end{aligned}$$

  This is as large as it can get - the maximal amount of new information we can get on X is quantified in its entropy, and we can learn all of it by looking at Y.

- Something in the middle: $p(0,0) = p/2, p(0,1) = p/2, p(1,1) = 1-p$.

$$\begin{aligned} H(X) &= -p\log p - (1-p)\log(1-p), \\ H(Y) &= -\frac{p}{2}\log\frac{p}{2} - (1-\frac{p}{2})\log(1-\frac{p}{2}), \\ H(XY) &= -\frac{p}{2}\log\frac{p}{2} - \frac{p}{2}\log\frac{p}{2} - (1-p)\log(1-p), \\ I(X;Y) &= -p\log p + \frac{p}{2}\log\frac{p}{2} - (1-\frac{p}{2})\log(1-\frac{p}{2}), \end{aligned}$$

  which is something in the middle.

Exercise 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

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

Solution

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

which is an equality only when knowing X tells us nothing about Y,that is, when Y, X are independent.