Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Quantum Information Theory
Tips 2

FS 11
TA Lídia del Rio

Exercise 2.1 Entropy as a measure of uncertainty

Let me try to introduce you to entropy measures in an intuitive way. For a precise and neat formulation you should read pages 13-17 of the script.
The goal here is to find a way of quantifying the ignorance one has about a given phenomenon. The first step is to express our knowledge with probability distributions. As usual we start with a rather ridiculous example: my father forgot where he left his glasses again, and gave my brother, my sister and me ten minutes to try to find them. Our house has 6 rooms: kitchen, parents bedroom, children bedroom, dinning room, living room and bathroom. Each of us searched for a while and tried to figure where the glasses could be. Our knowledge after ten minutes is represented in Fig. 1.

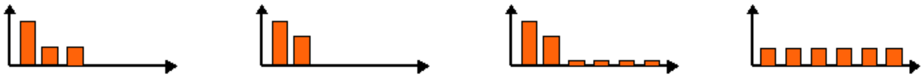

Figure 1: Our knowledge on the whereabouts of the glasses: I (first graph) searched in three of the rooms and found nothing so I think it is either in my dad's room (the most likely option, given our previous experience), in the kitchen or in the bathroom. My brother (second graph) also searched in the bathroom, narrowing his options to parents room and kitchen. My sister (third graph) is almost in the same conditions except that she did not search very thoroughly so assumes they may have escaped her in one of the other rooms. The cat (fourth graph) has no idea.

These probability distributions have some things in common (like the height of the peak probability for the three siblings, or the size of the support, for my sister and the cat), but also some differences — it is clear that we have different degrees of certainty about where the glasses may be.
So, looking at those probability distributions, how can we quantify our knowledge, or lack of it? The answer is: it depends. It depends on how you want to use your knowledge.
If your father was feeling lucky and just asked us what was the one room where he was more likely to find the glasses, i.e. if we had to guess where the glasses were in a single shot, than the relevant quantity would be the peak probability, that gives us the probability of guessing correctly. In this case one could use the min-entropy to quantify our ignorance,

$$H_{\min}(X)_P = -\log \max_{x\in\mathcal{X}} P_X(x). \tag{1}$$

In this case my guess is as good as my the one by my sister — even though our state of knowledge is clearly different. This is how the min-entropy sees these probability distributions:

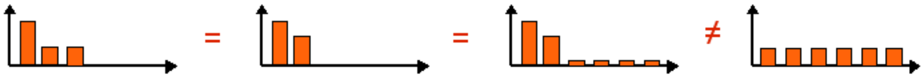

Figure 2: The min-entropy only cares about the peak probability, that expresses the probability of taking a correct guess in a single shot approach, ignoring how the other the probabilities are spread.

On the other hand, if our father wanted to be absolutely sure he would find the glasses and asked us to give him a list of the rooms where the glasses could still be, even if unlikely cases, the peak probabilities would be irrelevant and the only thing to matter would be the support of the distribution.

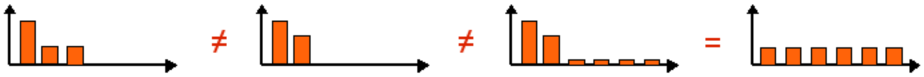

Figure 3: The max-entropy cares about the support of the distribution and is indifferent to the way the probabilities are spread. This makes it see similar distributions as if they were very different (for instance graphs 2 and 3), while two distributions that are quite distinct (like 3 and 4) may have the same max-entropy.

The max-entropy gives us the size (in bits) of the memory that is required to store all the possibilities, so it would be suited to quantify our ignorance in this case.

$$H_{\max}(X)_P = \log |P_X|. \tag{2}$$

You may think that it is somehow *unfair* that my sister, who knows much more than the cat about where the glasses may be, is considered as ignorant as it according to this entropy measure. In particular, you may find it a bit of a waste of time (or, in any another example, memory) that she sends our father to look for the glasses in the four rooms where she is *almost* certain they will not be. If she takes a very small chance of being wrong — a small error tolerance — she may dismiss the very unlikely event that the glasses are in one of those four rooms and tell the father to search only in the first two rooms.

Fortunately *someone* in the information theory community thought about this question before my sister and introduced the smooth max-entropy,

$$H^{\epsilon}_{\max}(X)_P = \min_{Q_X \in \mathcal{B}^{\epsilon}(P_X)} H_{\max}(X)_Q, \tag{3}$$

where the minimum is taken over all probability distributions Q_X that are ϵ-close to P_X according to the trace distance. In practice the smooth max-entropy takes the tail of the distribution and wipes unlikely events until it reaches a maximum weight of ϵ.

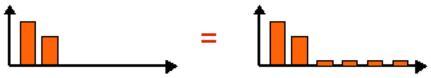

Figure 4: For the smooth max-entropy these two distributions look the same if the total probability of the four unlikely events in the tail of the second distribution sum up to less than the error tolerance ϵ .

There is also a smooth min-entropy,

$$H^{\epsilon}_{\min}(X)_P = \max_{Q_X \in \mathcal{B}^{\epsilon}(P_X)} H_{\min}(X)_Q, \tag{4}$$

which does not have an immediate one-shot meaning. We will see next week that in the iid limit the smooth min- and max-entropies converge.

Small remark before we continue: one question you may ask is "why are all these entropies logarithmic?". The short answer is that we want them to be additive, i.e. we want the entropy of two uncorrelated events to be the sum of the individual entropies, or in other words "what we do not know about two things that have nothing to do with each other is just the sum of our ignorance on each of the things" (what I do not know about oranges and war tanks is my ignorance about oranges plus... I am sure you got it by now). More in page 13 of the script.

My intuition for the Shannon entropy is not as clear as for the previous two (I am hoping these two were clear!). Let me try anyway. Sticking to the example, if our father had lost the glasses many times and we had reached the same probability distributions all those times, then the Shannon entropy would tell us how *surprised*, in average, we would be with the actual whereabouts of the glasses.

It makes sense that the smaller the probability of an event, the more surprised we would be if that event happened. Since we everything is logarithmic here, we define the surprise content, or *surprisal* of an event E as $h(E)_P = -\log P_X(E)$ (again, check page 13 of the script for more convincing arguments). The Shannon entropy is just the expectation value of the surprisal, i.e. its average over all possible outcomes,

$$H(X)_P = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x). \tag{5}$$

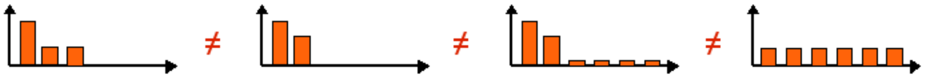

Figure 5: All probability distributions are different for the Shannon entropy. This may seem good but has a drawback: it has no operational meaning for a single shot experiment. In particular, very similar distributions, like those of graphs 2 and 3, may have very different Shannon entropies.

More remarks: In general $H_{\min}(X)_P \leq H(X)_P \leq H_{\max}(X)_P$. In the worst case scenario, i.e. the uniform distribution, they are all the same. In the i.i.d. limit their smooth versions are all the same too (but we will cover that later).

Exercise 2.2 Mutual Information

Two new things in this exercise: conditional entropies and mutual information.
Conditional entropy quantifies our ignorance about something given our knowledge about a (hopefully) related event — for instance our uncertainty about the weather tomorrow after listening to the radio forecast.
The Shannon conditional entropy of X given Y is defined as the expectation value of the surprisal of x knowing $Y = y$,

$$\begin{aligned} H(X|Y) &= \langle h(x|Y=y)_{P_{xy}} \rangle_{xy} \\ &= \langle -\log P_{X|Y=y}(x) \rangle_{xy} \\ &= -\sum_{x,y} P_{XY}(x,y)\ \log P_{X|Y=y}(x). \end{aligned} \tag{6}$$

As we have $P_{X|Y=y}(x) = P_{XY}(x,y)/P_Y(y)$, so comes

$$H(X|Y) = H(XY) - H(Y). \tag{7}$$

Conditional min- and max-entropies are given on page 16 of the script.
The mutual information tells us how correlated two experiments (read random variables) are. If they are maximally correlated (like a very accurate forecast and the actual weather) then you can determine each one of them from the other. If they are uncorrelated (like solar flames and the price of gold) then knowing one of them does not help you at all to guess the other.
The mutual information between X and Y is defined in a natural way as "what we have learned about X by knowing Y", or "what we know about X now that we know Y minus what we knew about X before knowing Y", or, to make it even more bizarre, "what we *did not know* about X before minus what we *do not know* about X now that we know Y", i.e. the entropy of X minus the conditional entropy of X given Y,

$$I(X:Y) = H(X) - H(X|Y). \tag{8}$$

Notice that the mutual information is symmetric, $I(X:Y) = H(X) + H(Y) - H(XY) = I(Y:X)$.
Back to the exercise, you have to apply this concept to calculate the mutual information between your guess and the actual weather, and also between the guess of your grandfather and the weather.
The conditional and marginal probabilities for the radio forecast case are represented in the figure below. Check solution sheet 1 for details. Remember that in the case of a sunny forecast any strategy was equally bad — so for simplicity you may assume you trust the radio report and say it will not rain.

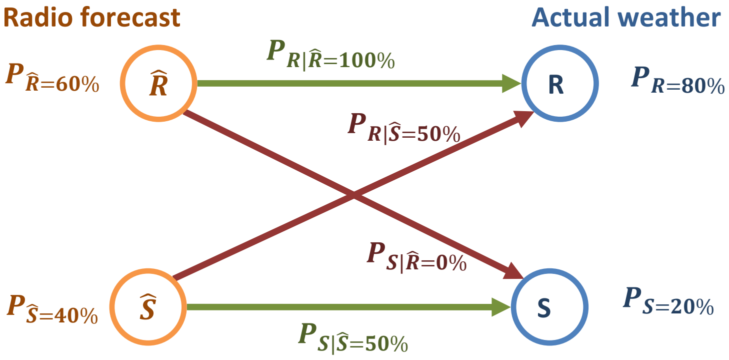

Figure 6: The radio forecast and the actual weather: marginal and conditional probabilities. Naturally, you can get the joint probabilities using $P_{X|Y=y}(x) = P_{XY}(x,y)/P_Y(y)$.

Exercise 2.3 Channel capacity

Channels! A channel is a rather intuitive concept. Think of a noisy telephone line from the thirties. The question here is: how do we characterise the telephone line? We want to know how well a person on the other side will understand us when we phone. The relevant parameters cannot be the input sounds — those will change each time we use the channel. We are more interested in how reliably the telephone will reproduce each sound input: each time I say “aye”, what is the probability that the sound that arrives the other side is “aye” and not “nay”? In other words, what is the probability of getting an “aye” *conditioned* on the fact that I input an “aye”? You can see where this is leading. A channel is fully characterised by the set of conditional probabilities of the outputs given each of the inputs. Pages 10–11 of the script have details and a much more precise formulation of what a channel is.

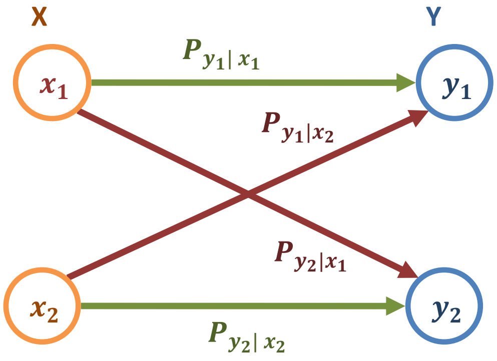

Figure 7: A channel with two inputs x_1 and x_2 and two outputs is defined by the conditional probabilities $P_{y_i|x_j}$.

You may see that in exercise 2.2 we had a channel — expect that in that case we also fixed the probabilities of each input. Now that you have characterised your telephone line with all the conditional probabilities, you want to find a way of quantifying how reliable it is. One way of doing this is to ask “I want to send a message through this channel with only a negligible probability of error. How long can that message be?” In the iid limit (i.e. you use the channel many times), the answer is the *capacity* of the channel. This is explained in detail in pages 18–22 of the script. Here as usual I will just try to give a feeling of its meaning.

You have seen that the mutual information gives us an amount of how correlated two things are. That is precisely what we want of a channel — the more correlated the input and output are, the better the channel. The quality (or capacity) of a channel should be related to the mutual information between input and output.

There is one free parameter in a channel, which is the probability distribution on the inputs. We can use it to maximise the certainty that our message will be well received by *encoding* our message. For instance, imagine a channel that transmits “ayes” correctly with 99% of probability but fails at transmitting “nays” 30% of the time. We may use redundancy to ensure our “nays” will be understood as such, by saying “nay nay nay” for each “nay” intended. The person in the other side will *decode* any sequence of two or three “‘nays” (and one or none “aye”) as a single “nay”.

So, as we can use P_X to maximise the fidelity of the channel, the final capacity is given by

$$C = \max_{P_X} I(X:Y). \tag{9}$$

In part *a*) you have to apply this to two simple channels. You will find that the distribution P_X that maximises the mutual information is the uniform distribution. In part *b*) you are going to prove that that is the case for all symmetric channels.

You start by considering N probability distributions for the input, $P_X^1, \dots P_X^N$, such that $I(X:Y)_{P^i} = I(X:Y)_{P^j}, \forall i, j$. As an example you can think of a symmetric channels, where a permutation of the input probability distribution does not change the mutual information between input and output: $P_X^2, \dots P_X^N$ could be permutations of P_X^1.

Now suppose that Joanna chooses which probability distribution she will use for an input by picking a ball from a bag at random. Formally, this is expressed by a random variable B that can take values $b = 1, \dots N$ (assume a uniform probability distribution on the outcomes of B).

Now you compare the mutual information between input and output of the channel for Joanna, who knows which ball she picked—and therefore which P_X^i she chose as input, $I(X:Y|B)$, and someone who does not know which distribution she

chose, $I(X:Y)$. Use properties of the conditional entropies to prove this; in particular, do not forget that knowing more cannot hurt ($H(A|B) \leq H(A)$), and that the conditional probabilities that define the channel are fixed.
You should get that $I(X:Y|B) \leq I(X:Y)$, i.e. one is always better if one does not know which P_x^i was used. "Not knowing which distribution was used" is the same as admitting that a uniform mixture of those distributions was used, i.e. P_X^1 with probability $1/N$, P_X^2 with probability $1/N$, etc. But what does that mean for symmetric channels? When all the P_x^i are permutations of each other, what is their uniform mixture? Up to you to work out!