ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Quantum Information Theory
Tips 2

HS 13
Aleksejs & David

Exercise 2.1 Information and Description Length

Try to solve this exercise without additional hints :-).

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 $Y = y$,

$$\begin{aligned} H(X|Y) &= \langle h(x|Y=y)_{P_{xy}} \rangle_{xy} \\ &= \langle -\log P_{X|Y=y}(x) \rangle_{xy} \\ &= -\sum_{x,y} P_{XY}(x,y)\ \log P_{X|Y=y}(x). \end{aligned} \tag{1}$$

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

$$H(X|Y) = H(XY) - H(Y). \tag{2}$$

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,

$$I(X:Y) = H(X) - H(X|Y). \tag{3}$$

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.

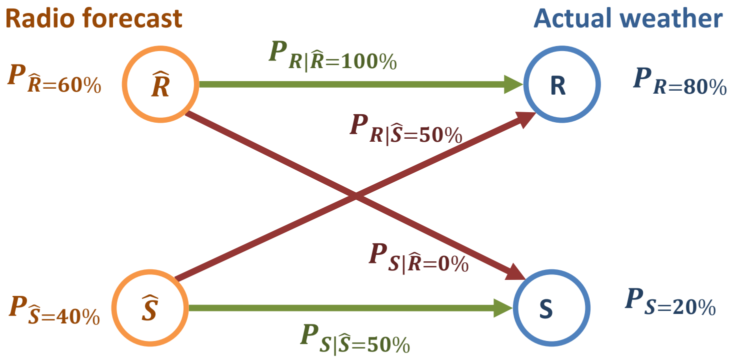

Figure 1: The radio forecast and the actual weather: marginal and conditional probabilities. Naturally, you can get the joint probabilities using $P_{X|Y=y}(x) = P_{XY}(x, y)/P_Y(y)$.

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.

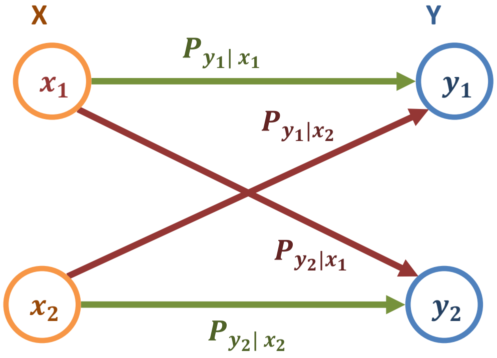

Figure 2: A channel with two inputs x_1 and x_2 and two outputs is defined by the conditional probabilities $P_{y_i|x_j}$.

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

$$C = \max_{P_X} I(X : Y). \tag{4}$$

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!