Exercise Sheet 1

Exercise 1.1.
Let $\mathcal{C}$ be a channel with matrix $P = (p(y|x))$, input alphabet Σ and output alphabet O. Recall the following definitions from the lecture.

A channel is called *input (output) symmetric* if there exist permutations μ and τ on Σ and O, respectively, with $\mu^q = \mathrm{id}_\Sigma$ and $\tau^q = \mathrm{id}_O$, where $q = |\Sigma|$, such that all orbits of μ (all orbits of τ) have q elements and such that for all $x \in \Sigma$ and $y \in O$ we have $p(y|x) = p(\tau(y)|\mu(x))$.

A channel is called *binary* if Σ has only two elements. A *binary symmetric* channel is a binary channel that is symmetric. In all these cases, the permutations μ and τ giving rise to the definitions are involutions, i.e., $\mu^2 = \mathrm{id}_\Sigma$ and $\tau^2 = \mathrm{id}_O$.

We define the *error probability* of an input symmetric binary channel as

$$\frac{1}{2}\int_O \min(p(\tau(y)|1), p(y|1))\mathrm{d}y,$$

where we assume that $1 \in \Sigma$.

1. Let $\Sigma = O = \{0,1\}$,, $\varepsilon \in [0,1]$, $p(0|0) = p(1|1) = 1-\varepsilon$ and $p(0|1) = p(1|0) = \varepsilon$. This channel is the *binary symmetric channel* with crossover probability ε, denoted by $\mathrm{BSC}(\varepsilon)$. Show that $\mathrm{BSC}(\varepsilon)$ is symmetric, that is, define the corresponding maps μ and τ. What is its error probability? What is its capacity?

2. Let $\Sigma = \{0,1\}$, $O = \{0,1,\mathrm{E}\}$, $0 \leq \varepsilon \leq 1$, and let

$$p(y|x) = \begin{cases} 1-\varepsilon & \text{if } y = x \\ \varepsilon & \text{if } y = \mathrm{E} \\ 0 & \text{otherwise.} \end{cases}$$

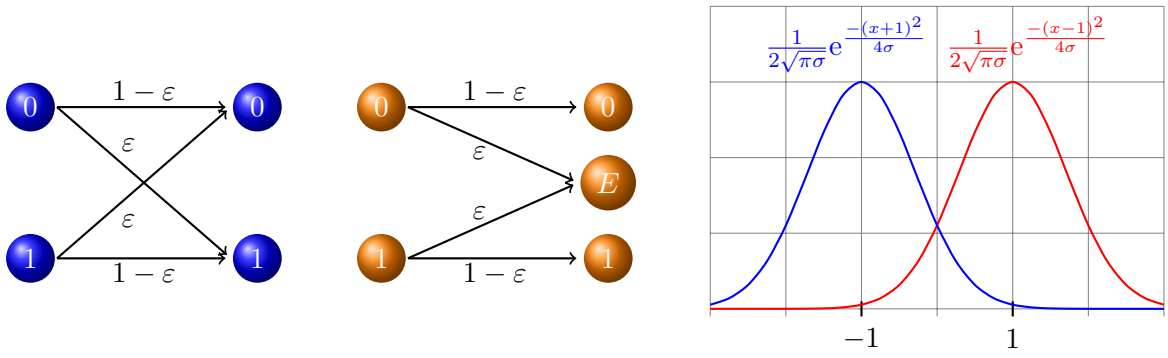
   This is the binary erasure channel with probability ε, denoted by $\mathrm{BEC}(\varepsilon)$. Show that $\mathrm{BEC}(\varepsilon)$ is an input symmetric channel. What is its error probability? What is its capacity?

3. Let $\Sigma = \{-1,+1\}$, $O = \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$, and $p(y,a) = \frac{1}{2\sqrt{\pi\sigma}}\mathrm{e}^{-(x-a)^2/4\sigma}$ for $a \in \{-1,+1\}$. This channel is the *binary input additive white Gaussian noise* channel with variance σ, and is denoted by $\mathrm{AWGN}(\sigma)$. Show that $\mathrm{AWGN}(\sigma)$ is symmetric. What is its error probability?

Exercise 1.2. Show that for a fixed length n, the Hamming distance is a metric on the space of the words of that length. Recall that a metric function satisfies the following conditions:

1. $d(x,y) \geq 0$ (non-negativity).

2. $d(x,y) = 0$ if and only if $x = y$ (identity of indiscernibles).



(a) (b) (c)

Figure 1: (a) The binary symmetric channel, (b) The binary erasure channel, (c) The binary input AWGN channel

3. $d(x, y) = d(y, x)$ (symmetry).
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

Exercise 1.3.

Consider the *q-ary symmetric channel* with probability ε. It is defined as follows: Let $|\Sigma| = q$, and $O := \Sigma$. Further, let $0 \leq \varepsilon \leq (q-1)/q$, and let $p(y|x) = 1 - \varepsilon$ if $y = x$, and $p(y|x) = \varepsilon/(q-1)$ if $y \neq x$.

For $y \in \Sigma^n$, show that any vector z of minimum Hamming distance to y maximizes $\prod_{i=1}^{n} p(y_i|z_i)$.

Exercise 1.4. In his seminal paper "A Mathematical Theory of Communication", Shannon justifies the definition of entropy as

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log p_i \tag{1}$$

by starting from a set of axioms that we expect any "measure of uncertainty" to satisfy, and proving that any function satisfying these axioms must be of the form given in equation (1), up to a multiplicative constant. More precisely, suppose we have a set of n events with probabilities $p_1, \ldots, p_n$. We are uncertain as to which event will occur, and we would like to define a measure $H(p_1, \ldots, p_n)$ of this uncertainty. We would like this measure to satisfy the following intuitive axioms:

1. H is continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H is a monotonic increasing function of n. Intuitively, this means that when all events are equally likely, there is more uncertainty when there are more events.
3. If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H. By breaking down a choice, we mean the

following: any subset of k events, say the k first events, can be grouped into one new event of probability

$$p'_1 = \sum_{i=1}^{k} p_i.$$

Then this new event can be decomposed again into k events of probability $\pi_1, \ldots, \pi_k$, with $\pi_i = \frac{p_i}{p_1+\cdots+p_k}$. Then choosing an event can be represented in two ways: on one hand, we can choose one of the original n events with probabilities $p_1, \ldots, p_n$, and on the other hand, we can choose one of the new $n-k+1$ events with probabilities $p'_1, p_{k+1}, \ldots, p_n$, and if the first event occurs, make another choice with probability vector $\pi_1, \ldots, \pi_k$. We would like the entropy to be the same in both cases, that is,

$$H(p_1, \ldots, p_n) = H(p'_1, p_{k+1}, \ldots, p_n) + p'_1 H(\pi_1, \ldots, \pi_k).$$

Note that the entropy term $H(\pi_1, \ldots, \pi_k)$ is weighted by p'_1 because this is the probability with which the second choice occurs.

We would like to prove that the only H satisfying the three axioms above is of the form

$$H(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i,$$

where K is a positive constant.

1. Let $A(n) := H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. Let s and t be integers. Using Axiom 3, show that $A(s^m) = mA(s)$. For n as large as we want, we can always find an m such that

   $$s^m \leq t^n < s^{m+1}.$$

   Using Axiom 2, deduce that

   $$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}$$

   and finally that $A(t) = K \log t$.

2. Suppose the p_i are commensurable probabilities, that is, they can be written as $p_i = \frac{n_i}{\sum_j n_j}$ (in other words, the ratio between any two probabilities is always a rational number). Using Axiom 3, prove that

   $$K \log \sum n_i = H(p_1, \ldots, p_n) + K \sum p_i \log n_i.$$

   Deduce that we have, in this case,

   $$H(p_1, \ldots, p_n) = -K \sum p_i \log p_i.$$

3. Now handle the incommensurable case by approximating the p_i by rational numbers and using Axiom 1.

Exercise 1.5.

Let $p(x)$ and $q(x)$ be two probability distributions. The *relative entropy* or *Kullback-Leibler distance* between p and q is defined as

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

The relative entropy is a measure of the "distance" between two distributions: if $p(x) = q(x)$ for every value x, then $D(p||q) = 0$.

Let X and Y be two random variables with joint distribution $p(x, y)$ and marginal distributions $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$. We define the *mutual information* $I(X; Y)$ as the relative entropy between the joint distribution $p(x, y)$ and the product of the marginals $p(x)p(y)$:

$$I(X; Y) = D(p(x, y)||p(x)p(y)) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

Intuitively, the mutual information is the cost of assuming that the variables X and Y are independent when they are not. Note that if X and Y are independent, $I(X; Y) = 0$.

1. Recall the definition of the joint entropy between X and Y as

   $$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

   and the conditional entropy $H(X|Y)$ as

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

   $H(Y|X)$ is defined similarly.

   Prove that

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

   From these properties we can view the mutual information, intuitively, as a measure of the amount of information that one random variable contains about another random variable.

   We can also deduce the *chain rule* for entropy:

   $$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

   This means that the uncertainty of the joint distribution of X and Y is identical to the uncertainty of one of the random variables plus the remaining uncertainty in the second when the first is known.

2. Let $X_1, \ldots, X_n$ be discrete random variables with joint distribution $p(x_1, \ldots, x_n)$. Their joint entropy is defined similarly as

$$H(X_1, \ldots, X_n) = -\sum_{x_1,\ldots,x_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n).$$

   Using the chain rule for two variables, prove that

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

3. Using the definition of relative entropy and of conditional probability, prove that

$$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)).$$

Exercise 1.6. Recall *Jensen's inequality*: For f a convex function and X a discrete random variable,

$$E[f(x)] \geq f(E(X)).$$

If f is strictly convex, this is true with equality only if $X = E[X]$ almost surely (i.e, X is a constant).

1. Applying Jensen's inequality to the logarithm function, which is strictly concave, show that

$$D(p||q) \geq 0, \tag{2}$$

   with equality if and only if $p(x) = q(x)$ for all x.

   Deduce that for any pair X, Y of random variables, we have

$$I(X, Y) \geq 0, \tag{3}$$

   with equality if and only if X and Y are independent.

2. Let X take its values in the set χ. Deduce from (2) that $H(X) \leq \log |\chi|$, with equality if and only if X is uniformly distributed over χ.
   Hint: consider $D(p||u)$, where u is the uniform distribution over χ.

3. Deduce from (3) that $H(X|Y) \leq H(X)$, and then that

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i),$$

   with equality if and only if the X_i are independent.