TUTORIAL II

1 Homework 2

1. Let $X \in \mathbb{N}$ be a discrete random variable and $g : \mathbb{N} \to \mathbb{N}$. What can you say in general on the relation between $H(X)$ and $H(g(X))$? And in particular, if $g(n) = 2^n$?
2. We know that more information cannot increase uncertainty in the sense that $H(X|Y) \leq H(X)$. Show that this is not true if we do not take the average of Y, i.e., give an example of a pair of random variables (X, Y) such that $H(X|Y = y) > H(X)$ for some y.
3. Show that $H(X|Y) = 0$ implies that X is a (deterministic) function of Y.
4. Suppose (X_i, Y_i) for $i \in \mathbb{N}$ are chosen iid according to the distribution P_{XY}. We write X^n for the sequence $(X_1, \ldots, X_n)$. What is the limit of the sequence of random variables $\frac{1}{n} \log \frac{P_{X^n}(X^n) \cdot P_{Y^n}(Y^n)}{P_{X^n Y^n}(X^n, Y^n)}$?
5. Find a distribution (p_1, p_2, p_3, p_4) on elements $\{1, 2, 3, 4\}$ such that there are two codes with different encoding lengths $\{\ell_i\}_{1 \leq i \leq 4}$ and $\{\ell'_i\}_{1 \leq i \leq 4}$ while both codes minimize the average length $\sum_i p_i \ell_i$.

2 Entropy of Markov chains

A *Markov chain* is an indexed sequence $\{X_i\}$ of random variables such that the variable X_{n+1} only depends on the value of X_n. In other terms:

$$\mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n)$$

In the following, we will always assume that the Markov chains are time-independant, ie the following holds:

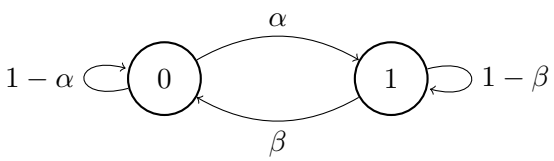
$$\mathbf{P}(X_{n+1} = a | X_n = b) = \mathbf{P}(X_1 = a | X_0 = b)$$

In this case, the evolution of the system depends only on the conditional distribution $P(X_1|X_0)$, and we will usually describe this distribution using a *probability transition matrix* $P = [P_{ij}]$, where $P_{ij} = \mathbf{P}(X_1 = j | X_0 = i)$. If all the X_i's can only take a finite number of value, we usually represent X_i by its distribution $p_i = (\mathbf{P}(X_i = 0), \mathbf{P}(X_i = 1), \ldots, \mathbf{P}(X_i = l))$.

Those notations allow us to use the tools of linear algebra, since we can describe the dependency between X_{i+1} and X_i using the matrix product: $p_{i+1} = p_i \cdot P = p_0 \cdot P^i$. For instance, under reasonable assumptions, we know that P^i converges to a certain matrix P^∞, and that the resulting limit distribution $p_\infty = p_0 \cdot P^\infty$ is the only fixpoint of P (i.e. the only p such that $p = p \cdot P$).

1. Find the stationary/limit distribution of a two-states Markov chain with a probability transition matrix of the form:

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$



2. In this case of a system with memory, the basic notion of entropy don't capture the dependency between states. Thus, we define another notion of entropy: the *entropy rate* is defined as

$$H(\mathcal{X}) = \lim_{n\to+\infty} H(X_n|X_{n-1},\dots,X_0) = \lim_{n\to+\infty} \frac{1}{n} H(X_1,\dots,X_n)$$

In the case of Markov chain, we thus have: $H(\mathcal{X}) = \lim_{n\to+\infty} H(X_n|X_{n-1})$. If we are in a convergent case, we have: $H(\mathcal{X}) = H(X_1|X_0)$, where the conditional entropy is calculated using the stationary distribution, ie with $X_0 \sim \mu$.
Compute the entropy rate of the Markov chain of question 1.

3. What is the maximum value of $H(\mathcal{X})$ in this example ?

4. We now take the special case where $\beta = 1$. Give a simplified expression of the entropy rate.

5. Find the maximum value of $H(\mathcal{X})$ in this case. Is it normal that this maximum is achieved for $\alpha < 1/2$?

6. Let $N(t)$ be the number of allowable state sequences of length t for the Markov chain (with $\beta = 1$). Find $N(t)$ and calculate:

$$H_0(\mathcal{X}) = \lim_{t\to+\infty} \frac{1}{t} H_0(X_0,\dots,X_{t-1}) = \lim_{t\to+\infty} \frac{1}{t} \log N(t)$$

Why is H_0 an upper bound on the entropy rate of the Markov chain ? Compare H_0 with the maximum entropy found in the previous question.

3 Code for unknown distribution

Recall that we can build a code C that achieve an expected description length $l(C)$ within 1 bit of the lower bound, that is:

$$H(X) \leq l(C) < H(X) + 1$$

This is done using the following choice of word lengths: $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$. In some case, we don't know the true distribution p, but only have an approximation q, and still want to find a code.

1. Show that if we use the same choice of word lengths: $l_i = \left\lceil \log \frac{1}{q_i} \right\rceil$, we have:

$$H(p) + D(p||q) \leq E_p(l(C)) < H(p) + D(p||q) + 1$$

4 From fair coins to any discrete distributions

Given a random variable X following a specific discrete distribution p, we want to know how many fair coins does it take to generate X. We want to minimize the average number of tosses we have to make.

More formally: we are given a sequence of fair tosses $Z_1, Z_2, \dots$, and wish to generate a discrete random variable $X \in \mathcal{X} = \{1,\dots,m\}$, with a distribution $p = (p_1,\dots,p_m)$. Let T be the random variable denoting the number of coins flips used in the algorithm.
We can describe the algorithm using a tree: the leaves are marked by output symbols X, and the path to the leaves is given by the sequence of bits produced by the fair coin. We moreover assume that the tree satisfies some properties:

- The tree should be complete (i.e. every node is either a leaf or has two descendants)
- The probability of a leaf at depth k is 2^{-k}. Many leaves may be labeled with the same output symbol – the total probability of all these leaves should be the one corresponding to this output symbol in the distribution p.

In this representation, the average number of tosses is the expected depth of the tree. We want to find a tree with such properties that minimize its expected depth.

1. Consider the following distribution for X:

$$X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & \text{with probability } \frac{1}{4} \\ c & \text{with probability } \frac{1}{4} \end{cases}$$

   Find the minimal average number of fair bits (tosses) needed to generate X. Compare this value with $H(X)$.

2. Given a complete tree, we denote by $\mathcal{Y}$ the set of the leaves. Consider a distribution Y on the leaves such that the probability of a leaf at depth k is 2^{-k}. Show that the expected depth of the tree is equal to the entropy of such a distribution.

3. Show that for any algorithm generating X, the expected number of fair bits used is greater than the entropy, i.e. that: $ET \geq H(X)$.

4. Show that if all the p_i's are dyadic (i.e. $p_i = 2^{-l_i}$), one can achieve $ET = H(X)$ with a finite algorithm.

5. Now we want to extend this result for non-diadic distributions. We will assume that this result holds even in the infinite case: i.e. for a dyadic distribution over an infinite set $\mathcal{Y}$, we still can find an (infinite) algorithm T that achieves $ET = H(Y)$.

   (a) Let's begin with an example: give an infinite tree that generate a random variable X with a distribution $(\frac{1}{3}, \frac{2}{3})$. What is its expected height ? Compare this value with $H(X)$.

   (b) Given a non-dyadic distribution $p = (p_1, \ldots, p_m)$, we split it into dyadic atoms, for example $p_1 \to (p_1^{(1)}, p_1^{(2)}, \ldots)$, and so on. We take the tree T that achieves $H(Y) = ET$, and want to show that it achieves the following inequalities:

$$H(X) \leq ET < H(X) + 2$$

   We already proved the first inequality in a previous question. Show that the second inequality is equivalent to $H(Y|X) < 2$.

   (c) Expanding the entropy of Y, we have:

$$H(Y) = -\sum_{i=1}^{m} \sum_{j \geq 1} p_i^{(j)} \log p_i^{(j)} = \sum_{i=1}^{m} \sum_{j : p_i^{(j)} > 0} j 2^{-j}$$

   For $i \in [1; m]$, we denote the corresponding term in the expansion by T_i, i.e.:

$$T_i = \sum_{j : p_i^{(j)} > 0} j 2^{-j}$$

   Show that in order to prove the upper bound, it's enough to prove that for all i, $T_i < -p_i \log p_i + 2p_i$.

   (d) Denote by n the only integer such that: $2^{-(n-1)} > p_i \geq 2^{-n}$, so we can rewrite $\sum_{j : p_i^{(j)} > 0} \ldots$ into $\sum_{j : j \geq n, p_i^{(j)} > 0} \ldots$. Using the fact that $p_i = \sum_{j : \ldots} p_i^{(j)}$, show that $T_i + p_i \log p_i - 2p_i < 0$. Conclude.