TUTORIAL II

0 Homework 1

1. (Repetition code) Suppose that you have a disk drive where each bit gets flipped with probability $f = 0.1$ in a year. In order to be able to correct errors, we take a copy of the full drive $N-1$ times so that we have N copies of the original data (N is odd). After one year, I would like to retrieve a given bit of the original drive. What should I do? Suppose I want the probability of error for this bit to be at most δ, how large should I take N as a function of δ? How large is this for $\delta = 10^{-10}$?

2. 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 $g(n) = 2^n$?

1 Axiomatic approach to the Shannon entropy

If we require certain properties of our uncertainty measure, then it uniquely specifies the Shannon entropy. Let $\Delta_m = \{(p_1, \ldots, p_m) \in \mathbb{R}^m : p_i \geq 0, \sum_i p_i = 1\}$ be the set of distributions on m elements. Let our uncertainty measure $H_m : \Delta_m \to \mathbb{R}$ be a sequence of functions satisfying the following desirable properties

1. Symmetry: For any $m \geq 1$ and any permutation π of $\{1, \ldots, m\}$, $H_m(p_1, \ldots, p_m) = H_m(p_{\pi(1)}, \ldots, p_{\pi(m)})$

2. Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = 1$

3. Continuity: For any $m \geq 1$, H_m is a continuous function

4. Grouping: For any $m \geq 2$,
$$H_m(p_1, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

5. Monotonicity: We have $H_m(\frac{1}{m}, \ldots, \frac{1}{m}) \leq H_{m+1}(\frac{1}{m+1}, \ldots, \frac{1}{m+1})$

Prove that $H_m(p_1, \ldots, p_m) = -\sum_{i=1}^m p_i \log_2 p_i$.

You can proceed in the following way. Let $g(m) = H_m(\frac{1}{m}, \ldots, \frac{1}{m})$.

1. Show that $g(n \cdot m) = g(n) + g(m)$.

2. Conclude that $g(m) = \log_2 m$. (Hint: for any n, let ℓ_n be such that $2^{\ell_n} \leq m^n \leq 2^{\ell_n + 1}$, show that $\frac{\ell_n}{n} \leq g(m) \leq \frac{\ell_n + 1}{n}$).

3. Use this to compute the value of $H_2(p, 1-p)$.

4. Conclude with H_m.

2 Data processing inequality for mutual information

Recall that:

$$H(X|Y) \stackrel{\text{def}}{=} \sum_{y \in A_Y} P_Y(y) H(X|Y=y) \quad , \quad H(X,Y) = H(X) + H(Y|X) \quad \text{and} \quad I(X;Y) \stackrel{\text{def}}{=} H(X) - H(X|Y)$$

0. 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.

We define the conditional mutual information:

$$I(X;Y|Z) \stackrel{\text{def}}{=} H(X|Z) - H(X|Y,Z)$$

If X and Z are conditionally independent given Y (i.e. $\mathbf{P}_{Z|Y,X} = \mathbf{P}_{Z|Y}$), we will use the notation $X \to Y \to Z$ (this notation is motivated by the theory of Markov chains). Notice that $X \to Y \to Z$ implies $Z \to Y \to X$ since $\mathbf{P}_{Z|Y,X} = \mathbf{P}_{Z|Y} \Rightarrow \mathbf{P}_{X|Y,Z} = \mathbf{P}_{X|Y}$.

1. Show that $I(X;Y|Z)$ is the average over Z of $I(X;Y)$, ie: $I(X;Y|Z) = \sum_z \mathbf{P}(Z = z) I(X|Z = z; Y|Z = z)$.

2. Show that $I(X;(Y,Z)) = I(X;Z) + I(X;Y|Z)$

3. For any $X \to Y \to Z$, show that the conditional mutual information $I(X;Z|Y)$ is 0.

4. Using question 2 and 3, show the data processing inequality: $I(X;Y) \geq I(X;Z)$ for any $X \to Y \to Z$.

5. Show that for any function g, we have $I(X;Y) \geq I(X;g(Y))$.

3 Code for unknown distribution

Recall that we can build a code C that achieves an expected length within 1 bit of the lower bound, that is:

$$H(X) \leq \mathbb{E}(|C(X)|) < H(X) + 1$$

This is done either by using Huffman's algorithm or using the following choice of word lengths: $l_x = \left\lceil \log \frac{1}{p(x)} \right\rceil$, where p is the distribution of X. In some cases, 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:

$$H(p) + D(p\|q) \leq \mathbb{E}(|C(X)|) < H(p) + D(p\|q) + 1$$

   Extra for those who know Huffman's algorithm: What about Huffman's algorithm?

4 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:

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

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

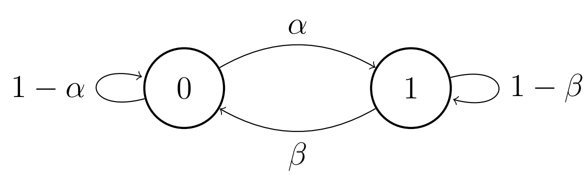
$$\mathbf{P}(X_{n+1} = a | X_n = b) = \mathbf{P}(X_1 = a | X_0 = b)$$

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 values, we usually represent X_i by its distribution $p_i = (\mathbf{P}(X_i = 0), \mathbf{P}(X_i = 1), \dots, \mathbf{P}(X_i = l))$.

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 the 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.