TUTORIAL II

1 Shannon's source coding theorem

Recall the definition of H_0^δ and of the typical sets:

$$T_\beta^N = \left\{ s \in \mathcal{X}^N : \left| \frac{1}{N} \log \frac{1}{\mathbf{P}(s)} - H \right| < \beta \right\}$$

Theorem 1.1 (Source coding)**.** *For any δ, ε, there exists N_0 such that, for $N \geq N_0$, we have:*

$$\left| \frac{1}{N} H_0^\delta(X^N) - H \right| < \varepsilon$$

Part 1: $\frac{1}{N} H_0^\delta(X^N) < H + \varepsilon$
The weak law of large numbers on the typical set T_β^N gives the following warranty:

$$\mathbf{P}(s \in T_\beta^N) \geq 1 - \frac{\sigma^2}{\beta^2 N}$$

We can compute the size of T_β^N using the fact that $s \in T_\beta^N \Leftrightarrow 2^{-n(H+\beta)} < \mathbf{P}(s) < 2^{-N(H-\beta)}$. Thus, since the total probability that T_β^N contains can't be bigger than 1, we have:

$$|T_\beta^N| \cdot 2^{-N(H+\beta)} < 1 \quad \Rightarrow \quad |T_\beta^N| < 2^{N(H+\beta)}$$

We take $\beta = \varepsilon$ and N large enough to have $\frac{\sigma^2}{\beta^2 N} < \delta$, and this gives $|S_\delta(X^N)| < |T_\varepsilon^N|$ (for N large enough). Taking the log gives the wanted inequality: $H_0^\delta(X^N) < N(H + \varepsilon)$.

Part 2: $\frac{1}{N} H_0^\delta(X^N) > H - \varepsilon$
We will do this direction assuming that this inequality doesn't holds: for any N, the smallest subset $S_\delta(X^N)$ is smaller that $2^{N(H-\varepsilon)}$. We will use the typical sets again to show that this is impossible: we take $\beta = \varepsilon/2$ and our task is to prove that a subset S' of size $|S'| < 2^{N(H-2\beta)}$ can't achieve $\mathbf{P}(x \in S') > 1 - \delta$ (for N larger than an N_0 we will determine).

1. Given a rival smaller subet S', we compute the probability of the subset using T_β^N:

$$\mathbf{P}(x \in S') = \mathbf{P}(x \in S' \cap T_\beta^N) + \mathbf{P}(x \in S' \cap \overline{T_\beta^N})$$

   Give a majoration of both term using hypothesis and Part 1 of the proof.

2. Put the two terms together and conclude this part of the proof.

2 A more realistic find query

We consider a list of 32 elements and we want to test if a given element z belong to the list or not. We assume that all the probability that the element belongs to the list is $1/2$, and that all the position within the list are equiprobable. Our strategy is to test the first element, then the second element, ... until the wanted element is found or the end of the list is reached. We denote by F the random variable which is equal to 1 if and only if z is in the list, 0 otherwise.

1. Compute the entropy of F.
2. We denote by L_1 the random variable corresponding to the result of the first test. Compute the entropy of L_1.
3. Compute the distribution of the joint variable (F, L_1), and give the joint entropy $H(F, L_1)$.
4. Compute the conditional entropy $H(F|L_1)$.
5. We denote by $L_2, \ldots, L_n$ the result of the successive tests. Compute directly the conditional entropy $H(F|L_1, \ldots, L_n)$.
6. If we plug $n = 16$ in the previous solution, we find $0.689 > \frac{1}{2}$. Is it reasonable ? What is the value of $H(F|L_1, \ldots, L_{32})$?

3 Data processing inequality for mutual information

Recall that:

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

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

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).

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

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:

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

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.