TUTORIAL III

0 Homework 2

1. Show that $H(X|Y) = 0$ implies that X is a (deterministic) function of Y.
2. We showed in class that the optimal $H(X) - \log_2(1 + \lfloor \log_2 |\mathcal{X}| \rfloor) \leq \mathbf{E}\{|C^*(X)|\} \leq H(X)$. Show that there is a distribution P_X such that the lower bound holds with equality. (We want a nontrivial example, i.e., $|\mathcal{X}| > 1$.)
3. Huffman's algorithm constructs a prefix code C_{H} given a distribution $(p_1, \ldots, p_m)$ on the symbols $\{1, \ldots, m\}$. The objective of this problem is to show that the expected length $L(C_{\mathrm{H}})$ is minimum among all the prefix codes. Huffman's algorithm constructs a binary tree as follows. The algorithm starts with independent nodes labeled by the elements $1, \ldots, m$ and the corresponding probability. At the beginning, all the nodes or marked unvisited. At each step, we choose the two unvisited nodes u, v with minimum value of p_u, p_v. We create a new node w with an assigned probability $p_w = p_u + p_v$ which is the parent of u and v. w is marked as unvisited and u, v are marked as visited. The step is repeated $m - 1$ times until we have one unvisited node (the root) with an assigned probability 1. To every path from the root to a leaf of the tree, we assign a bitstring where a "left" edge is read as 0 and a "right" edge is read as 1. The obtained tree defines a code in the following way: for any $x \in \{1, \ldots, m\}$, $C_{\mathrm{H}}(x)$ is the bitstring corresponding to the path from the root to x.
   (a) Show that for any optimal code, it can be transformed to one with the following property: the two longest codewords correspond to the two least likely symbols, and they have the same length and they only differ in the last bit.
   (b) Conclude that C_{H} achieves the optimal expected length for $(p_1, \ldots, p_m)$.
4. 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 $\sum_i p_i \ell_i$.

1 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?

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:

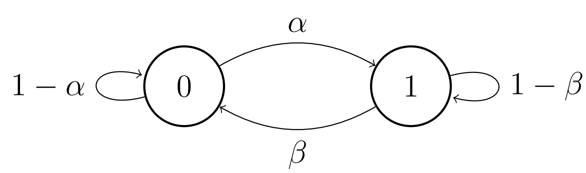
$$\mathbf{P}(X_{n+1} = a | X_n = b) = \mathbf{P}(X_1 = a | X_0 = b)$$

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 values, 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 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

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

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:

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

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 Fixed-length almost lossless compressor: source coding theorem

Recall that a *fixed length compressor* for source $Y \in \mathcal{Y}$ of length ℓ is a function $C : \mathcal{Y} \to \{0,1\}^\ell$. It has error probability at most δ if there exists a decompressor $D : \{0,1\}^\ell \to \mathcal{Y}$ such that $\mathbb{P}[D(C(Y)) = Y] \geq 1 - \delta$. Let define

$$\ell^{opt}(Y, \delta) = \min\{\ell : \text{ there exists a length } \ell \text{ compressor for } Y \text{ with error probability } \delta\}.$$

We will prove what is usually called *Shannon Source Coding Theorem*: Let $X^n = X_1 \dots X_n$ be a sequence of independent and distributed as $X \in \mathcal{X}$. For any $\delta \in (0,1)$,

$$\lim_{n\to\infty} \frac{\ell^{opt}(X^n, \delta)}{n} = H(X). \tag{1}$$

We will first show the lower bound (also called converse) and then give another proof for the upper bound.

3.1 Converse of Shannon source coding theorem

In this section, we will show that for any $\varepsilon > 0$ and for large enough n, we have $H(X) - 2\varepsilon \leq \ell^{opt}(X^n, \delta)/n$.

1. Prove that we are done if we can show that: for every set $S \subseteq \mathcal{X}^n$ such that $\mathbb{P}[X^n \in S] \geq 1 - \delta$ we have $|S| \geq 2^{n(H(X)-2\varepsilon)}$.

2. Now suppose that there is $S \subseteq \mathcal{X}^n$ such that $|S| \leq 2^{\ell n}$ for some ℓ and $\mathbb{P}[X^n \in S] \geq 1 - \delta$. Prove that

$$\mathbb{P}[X^n \in S] \leq \mathbb{P}_{X^n}\left[-\sum_{i=1}^{n} \log \mathbb{P}[X_i] \leq \ell n + \varepsilon n \right] + 2^{-\varepsilon n}.$$

3. Deduce that under the assumption of Question 2, $\ell \geq H(X) - 2\varepsilon$, and so the lower bound holds.

3.2 Achievability using random coding

Recall that in order to prove achievability of the source coding theorem, we chose the set S of correctly encoded symbols to be the set of $x^n \in \mathcal{X}^n$ such that $P_{X^n}(x^n) \geq 2^{-n(H(X)-\varepsilon)}$. We will now show a similar result by choosing the set S at random. In fact, we start by considering a general source (i.e., not necessarily iid) and derive an upper bound on the probability of error and will give us the desired result in the special case of an iid source.

Our objective is to show that for any source X and any integer $l \geq 0$, there exists a compressor with error probability

$$\delta \leq \mathbb{P}[-\log_2(P_X(X)) > l - \tau] + 2^{-\tau}, \;\; \forall \tau > 0. \tag{2}$$

1. Let $\tau > 0$, X be a random variable and C be a length l compressor. Let x_0 be a fixed letter of $\mathcal{X}$. Define $D = \{0,1\}^l \to \mathcal{X}$ by

$$D(y) = \begin{cases} x, & \text{if } \exists! x \in \mathcal{X} \text{ s.t. } C(x) = y \text{ and } -\log_2(P_X(x)) \leq l - \tau \\ x_0, & \text{otherwise} \end{cases} \tag{3}$$

Define also

$$J(x, C) = \{x' \in \mathcal{X} : C(x) = C(x'), x \neq x', \text{ and } -\log_2(P_X(x')) \leq l - \tau\}$$

Show that

$$\mathbb{P}[D(C(X)) \neq X] \leq \mathbb{P}[-\log_2(P_X(X)) > l - \tau] + \mathbb{P}[J(X, C) \neq \emptyset]$$

2. Let C be a random length-l compressor, that is for each $x \in \mathcal{X}$, $C(x)$ is a random bit string of length l, with each bit choosen independantly and uniformly from $\{0, 1\}$. Show that

$$\mathbb{E}_C[\mathbb{P}[J(X, C) \neq \emptyset]] \leq 2^{-\tau}$$

where we compute the mean on the randomness of C but not on X.

3. Prove Eq. (**??**)

4. Can you build from the proof a length l compressor with error $\delta \leq \mathbb{P}[-\log_2(P_X(X)) > l - \tau] + 2^{-\tau}$?

5. Use Eq. (**??**) to give a proof of the upper bound in Shannon source coding theorem (**??**).