EE/Ma 126a Information Theory - Homework Set #6

Ling Li, ling@cs.caltech.edu

November 15, 2000

Remark: All references go to chapter 4 of the textbook *Elements of Information Theory.*

6.1 *Stationary processes.*

(a) $\boxed{H(X_n|X_0) = H(X_{-n}|X_0) \text{ is true}}$. The stationarity gives

$$\Pr\{X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\} = \Pr\{X_{-n} = x_0, X_{-n+1} = x_1, \ldots, X_0 = x_n\}. \quad (1)$$

The entropy cares only the probability distribution, so $H(X_0, X_n) = H(X_{-n}, X_0)$.[1] Thus

$$\boxed{H(X_n|X_0) = H(X_0, X_n) - H(X_0) = H(X_{-n}, X_0) - H(X_0) = H(X_{-n}|X_0)}.$$

(b) $\boxed{H(X_n|X_0) \geq H(X_{n-1}|X_0) \text{ is not always true}}$. Consider a second-order Markov process $X_i = 1 - X_{i-2}$ where $X_i \in \{0, 1\}$. With the joint probability distribution of X_0 and X_1 as below,

$X_0 \backslash X_1$	0	1
0	1/4	1/4
1	1/4	1/4

$\Rightarrow$

$X_1 \backslash X_2$	0	1
0	1/4	1/4
1	1/4	1/4

we can get the joint probability distribution of X_1 and X_2. Now we have

$$\Pr(X_0 = x_0) = \Pr(X_1 = x_0) = \Pr(X_2 = x_0) = \frac{1}{2}$$

and

$$\Pr(X_0 = x_0, X_1 = x_1) = \Pr(X_1 = x_0, X_2 = x_1) = \frac{1}{4}.$$

These two equations ensure that this second-order (time invariant) Markov process is also a stationary process. Simple calculation gives

$$H(X_1|X_0) = \frac{1}{2}H(X_1|X_0 = 0) + \frac{1}{2}H(X_1|X_0 = 1) = 1$$

[1]This can also be strictly deduced. From (1), we get

$$\begin{aligned}
\Pr\{X_0 = x_0, X_n = x_n\} &= \sum_{x_1^{n-1}} \Pr\{X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n\} \\
&= \sum_{x_1^{n-1}} \Pr\{X_{-n} = x_0, X_{-n+1} = x_1, \ldots, X_{-1} = x_{n-1}, X_0 = x_n\} \\
&= \Pr\{X_{-n} = x_0, X_0 = x_n\}.
\end{aligned}$$

Thus $H(X_0, X_n) = H(X_{-n}, X_0)$.

and $H(X_2|X_0) = 0$ since X_2 is a function of X_0. Thus $\boxed{H(X_n|X_0) < H(X_{n-1}|X_0)}$ for $n = 2$ in this stationary process.

(c) $\boxed{\text{“}H(X_n|X_1^{n-1}, X_{n+1}) \text{ is non-increasing in } n\text{” is true}}$. Using the way we thought about the entropy and the stationary probability distribution in (a), we have $H(X_n|X_1^{n-1}, X_{n+1}) = H(X_{n+1}|X_2^n, X_{n+2})$.[2] Since conditioning reduces entropy, we have

$$\boxed{H(X_{n+1}|X_1^n, X_{n+2}) \leq H(X_{n+1}|X_2^n, X_{n+2}) = H(X_n|X_1^{n-1}, X_{n+1})}.$$

6.2 *The entropy rate of a dog looking for a bone.* For $i \geq 1$, the two consecutive positions X_{i-1} and X_i give the direction the dog takes at step i. Since the direction of step $(i+1)$ only depends on that of step i with probability $p = 0.1$ to change the direction, we know that X_{i+1} only depends on X_{i-1} and X_i, and

$$H(X_{i+1}|X_{i-1}, X_i) = H(p) = H(0.1) \approx 0.468996.$$

(a) The first step is equally likely to be positive or negative, so $H(X_1) = 1$. Thus

$$\boxed{H(X_1, X_2, \ldots, X_n) = H(X_1) + \sum_{i=1}^{n-1} H(X_{i+1}|X_{i-1}, X_i) = 1 + (n-1)H(0.1)}.$$

(b) The entropy rate of the dog walking is

$$\boxed{H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} \frac{1 + (n-1)H(p)}{n} = H(0.1) \approx 0.468996}.$$

(c) The probability that the dog takes i steps before reversing direction is $p(1-p)^{i-1}$. Thus the expected number of steps the dog takes before reversing direction is

$$\boxed{\sum_{i=1}^{\infty} i \cdot p(1-p)^{i-1} = \frac{1}{p} = 10}.$$

6.3 *Random walk on chessboard.* The Markov process associated with a random walk of a king on the chessboard is irreducible since the king can walk from any position to any position,

1	2	3
4	5	6
7	8	9

and aperiodic since the king can walk along $1 \to 2 \to 1$ or $1 \to 2 \to 4 \to 1$ (thus $d(1) = 1$). From the *Remark* at the end of section 4.2, the entropy rate $\mathcal{H}$ of this Markov process can be calculated

[2] Another way to get the equation: similar to 6.1(a), we get $H(X_1^{n-1}, X_{n+1}) = H(X_2^n, X_{n+2})$. Then

$$\begin{aligned} H(X_n|X_1^{n-1}, X_{n+1}) &= H(X_1^{n+1}) - H(X_1^{n-1}, X_{n+1}) \\ &= H(X_2^{n+2}) - H(X_2^n, X_{n+2}) = H(X_{n+1}|X_2^n, X_{n+2}) \end{aligned}$$

by assuming it is stationary. And from section 4.3 and *Example* 4.3.1, $\mathcal{H}$ can be calculated by (4.41), which can be further simplified as

$$\begin{aligned}
\mathcal{H} &= \log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \dots, \frac{E_m}{2E}\right) \\
&= \log(2E) + \sum_{i=1}^{m} \frac{E_i}{2E} \log \frac{E_i}{2E} = \frac{1}{2E} \sum_{i=1}^{m} E_i \log E_i, \qquad (2)
\end{aligned}$$

where E_i is the number of edges emanating from node i and $2E$ is the sum of E_i's. (2) also holds for Markov processes with a rook, bishops or a queen.[3]

Here I list the E_i's and $2E$ for each case in the chessboards.

$$\begin{array}{cccccc}
 & \text{king} & \text{rook} & \text{bishop 1} & \text{bishop 2} & \text{queen} \\
E_i = & \begin{array}{|c|c|c|}\hline 3&5&3\\\hline 5&8&5\\\hline 3&5&3\\\hline\end{array} & \begin{array}{|c|c|c|}\hline 4&4&4\\\hline 4&4&4\\\hline 4&4&4\\\hline\end{array} & \begin{array}{|c|c|c|}\hline 2& &2\\\hline &4& \\\hline 2& &2\\\hline\end{array} & \begin{array}{|c|c|c|}\hline &2& \\\hline 2& &2\\\hline &2& \\\hline\end{array} & \begin{array}{|c|c|c|}\hline 6&6&6\\\hline 6&8&6\\\hline 6&6&6\\\hline\end{array} \\
2E = & 40 & 36 & 12 & 8 & 56
\end{array}$$

Thus, the entropy rates are

$$\boxed{\begin{array}{rcll}
\mathcal{H}_{\text{king}} &=& \frac{1}{40}(4 \times 3 \log 3 + 4 \times 5 \log 5 + 8 \log 8) = \frac{1}{10}(3 \log 3 + 5 \log 5 + 6) & \approx 2.23645 \\
\mathcal{H}_{\text{rook}} &=& \frac{1}{36}(9 \times 4 \log 4) & = 2 \\
\mathcal{H}_{\text{bishop}_1} &=& \frac{1}{12}(4 \times 2 \log 2 + 4 \log 4) = \frac{4}{3} & \approx 1.33333 \\
\mathcal{H}_{\text{bishop}_2} &=& \frac{1}{8}(4 \times 2 \log 2) & = 1 \\
\mathcal{H}_{\text{queen}} &=& \frac{1}{56}(8 \times 6 \log 6 + 8 \log 8) = \frac{1}{7}(6 \log 6 + 3) & \approx 2.64425
\end{array}}$$

6.4 *Simple optimum compression of a Markov source.* Since the conditional probability distributions given S_i are all 2-*adic*, the codes designed for each state can all be optimal. One design of codes can be:

$U_{n-1} \backslash U_n$	S_1	S_2	S_3	L_i
S_1	0	10	11	1.5
S_2	10	0	11	1.5
S_3	n/a	0	1	1

The average message length L_i of the next symbol conditioned on the previous state S_i is also listed in the table above.

Easy to see that this Markov process is irreducible and aperiodic.[4] Thus from the *Remark* at the end of section 4.2, any initial distribution tends to the stationary distribution as $n \to \infty$. We can use the stationary distribution, which is

$$(\mu_1, \mu_2, \mu_3) = \left(\frac{2}{9}, \frac{4}{9}, \frac{3}{9}\right)$$

[3]The situation for a rook or queen is similar to that for a king. However, for certen type of bishop, it could not access part of the chessboard. Since the rules of chess forbide those accesses, we can remove the states related to those accesses and the entropy rate still can be calculated by (2).

[4]The 3 states S_1, S_2 and S_3 are in the same recurrent class and since the period of S_1 is 1, this class is aperiodic. So this class is ergodic, which is called ‘irreducible and aperiodic’ in the textbook.

here, to calculate the unconditional average number of bits per source symbol:

$$\boxed{L = \sum_{i=1}^{3} \mu_i L_i = 1.5 \times \frac{2}{9} + 1.5 \times \frac{4}{9} + \frac{3}{9} = \frac{4}{3}}.$$

And the entropy rate $\mathcal{H}$ can be calculated as (4.26), i.e.,

$$\boxed{\mathcal{H} = \frac{2}{9} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) + \frac{4}{9} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) + \frac{3}{9} H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{4}{3}}.$$

Thus we get $\boxed{L = \mathcal{H}}$.

6.5 *Shannon code.* Suppose all the probabilities are positive. For $i = 1, 2, \ldots, m$, we have

$$0 \leq F_i = \sum_{k=1}^{i-1} p_i < \sum_{k=1}^{m} p_i = 1.$$

Thus we can round off F_i to l_i bits (denoted by $\check{F}_i = \lfloor F_i \rfloor_{l_i}$) and use the bits of $\check{F}_i$ (which also are the first l_i bits of F_i) as a codeword for $X = i$. By the definition of round off, we have

$$F_i - 2^{-l_i} < \check{F}_i \leq F_i. \tag{3}$$

Also from $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$, we have

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1 \tag{4}$$

and

$$2^{-l_i} \leq p_i < 2^{1-l_i}. \tag{5}$$

(a) Since the probabilities are ordered so that $p_i \geq p_{i+1}$, we have $l_i \leq l_{i+1}$, hinting that the codeword for $X = i$ gets longer when i increases. From (3) and (5), for $1 \leq i < m$,

$$\check{F}_{i+1} + 2^{-l_{i+1}} > F_{i+1} = F_i + p_i \geq \check{F}_i + 2^{-l_i}.$$

So

$$\check{F}_{i+1} > \check{F}_i + 2^{-l_i} - 2^{-l_{i+1}} \geq \check{F}_i, \tag{6}$$

which means that all the codewords are distinct. For distinct codewords, one code cannot be the prefix of another code with the same or shorter length. So the codeword for $X = i$ cannot be the prefix of the codeword for $X = j$, if $i > j$.

When $i < j$, using (6) for $(j - i)$ times, we get $\check{F}_j > \check{F}_i + 2^{-l_i} - 2^{-l_j}$. Remember that $\check{F}_i$ and $\check{F}_j$ are l_i and l_j bits long respectively, and $l_i \leq l_j$. Thus

$$\check{F}_i + 2^{-l_i} - 2^{-l_j} = \check{F}_i + 2^{-l_i-1} + 2^{-l_i-2} + \cdots + 2^{-l_j} = \overbrace{\underbrace{0.xx\ldots x}_{l_i \text{ bits}}}^{=\check{F}_i} \underbrace{11\ldots1}_{(l_j - l_i) \text{ bits}}$$

is the biggest l_j-bit fraction that has $\check{F}_i$ as a prefix. Since $\check{F}_j$ is bigger than that, $\check{F}_i$ cannot be the prefix of $\check{F}_j$.

Thus for $i \neq j$, the codeword for $X = i$ cannot be the prefix of the codeword for $X = j$. That is, the code is prefix-free. $\square$

Summing (4) over $i = 1, 2, \ldots m$, we have

$$\sum_{i=1}^{m} p_i \log \frac{1}{p_i} \leq \sum_{i=1}^{m} p_i l_i < \sum_{i=1}^{m} p_i \left(\log \frac{1}{p_i} + 1 \right) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i} + 1.$$

That is

$$\boxed{H(X) \leq L = \sum_{i=1}^{m} p_i l_i < H(X) + 1}.$$

(b) For the probability distribution $(0.5, 0.25, 0.125, 0.125)$, we have $F_1 = 0$, $F_2 = 0.5 = (0.1)_2$, $F_3 = 0.75 = (0.11)_2$, $F_4 = 0.875 = (0.111)_2$, and $l_1 = 1$, $l_2 = 2$, $l_3 = l_4 = 3$. Thus the code is $\boxed{\{0, 10, 110, 111\}}$, which is apparently prefix-free.