16:332:542
Information Theory and Coding
Problem Set 2 Solutions

Chapter 2: 32, 33, 34
Chapter 3: 1, 3, 5
Chapter 4: 1, 2, 4, 6, 7, 10

Chapter 2

32. *Conditional entropy.* Under what conditions does $H(X \mid g(Y)) = H(X \mid Y)$?

Solution: (*Conditional Entropy*). If $H(X|g(Y)) = H(X|Y)$, then $H(X)-H(X|g(Y)) = H(X) - H(X|Y)$, i.e., $I(X;g(Y)) = I(X;Y)$. This is the condition for equality in the data processing inequality. From the derivation of the inequality, we have equality iff $X \rightarrow g(Y) \rightarrow Y$ forms a Markov chain. Hence $H(X|g(Y)) = H(X|Y)$ iff $X \rightarrow g(Y) \rightarrow Y$. This condition includes many special cases, such as g being one-to-one, and X and Y being independent. However, these two special cases do not exhaust all the possibilities.

33. *Fano's inequality.* Let $\Pr(X = i) = p_i, i = 1, 2, \ldots, m$ and let $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. The minimal probability of error predictor of X is $\hat{X} = 1$, with resulting probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$ to find a bound on P_e in terms of H. This is Fano's inequality in the absence of conditioning.

Solution: (*Fano's Inequality.*) The minimal probability of error predictor when there is no information is $\hat{X} = 1$, the most probable value of X. The probability of error in this case is $P_e = 1 - p_1$. Hence if we fix P_e, we fix p_1. We maximize the entropy of X for a given P_e to obtain an upper bound on the entropy for a given P_e. The entropy,

$$\begin{aligned} H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i & (2.120) \\ &= -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e & (2.121) \\ &= H(P_e) + P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right) & (2.122) \\ &\leq H(P_e) + P_e \log(m-1), & (2.123) \end{aligned}$$

since the maximum of $H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right)$ is attained by an uniform distribution. Hence any X that can be predicted with a probability of error P_e must satisfy

$$H(X) \leq H(P_e) + P_e \log(m-1), \tag{2.124}$$

which is the unconditional form of Fano's inequality. We can weaken this inequality to obtain an explicit lower bound for P_e,

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}. \tag{2.125}$$

34. *Monotonic convergence of the empirical distribution.* Let $\hat{p}_n$ denote the empirical probability mass function corresponding to $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x)$, $x \in \mathcal{X}$. Specifically,

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x) \tag{2.126}$$

is the proportion of times that $X_i = x$ in the first n samples, where I is an indicator function.

(a) Show for $\mathcal{X}$ binary that

$$ED(\hat{p}_{2n} || p) \leq ED(\hat{p}_n || p). \tag{2.127}$$

Thus the expected relative entropy "distance" from the empirical distribution to the true distribution decreases with sample size. *Hint:* Write $\hat{p}_{2n} = \frac{1}{2}\hat{p}_n + \frac{1}{2}\hat{p}'_n$ and use the convexity of D.

(b) Show for an arbitrary discrete $\mathcal{X}$ that

$$ED(\hat{p}_n||p) \leq ED(\hat{p}_{n-1}||p). \tag{2.128}$$

Solution: *Monotonic convergence of the empirical distribution.*

(a) Note that,

$$\begin{aligned}
\hat{p}_{2n}(x) &= \frac{1}{2n}\sum_{i=1}^{2n} I(X_i = x) \\
&= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} I(X_i = x) + \frac{1}{2}\frac{1}{n}\sum_{i=n+1}^{2n} I(X_i = x) \\
&= \frac{1}{2}\hat{p}_n(x) + \frac{1}{2}\hat{p}'_n(x).
\end{aligned}$$

Using convexity of $D(p||q)$ we have that,

$$\begin{aligned}
D(\hat{p}_{2n}||p) &= D(\frac{1}{2}\hat{p}_n + \frac{1}{2}\hat{p}'_n||\frac{1}{2}p + \frac{1}{2}p) \\
&\leq \frac{1}{2}D(\hat{p}_n||p) + \frac{1}{2}D(\hat{p}'_n||p).
\end{aligned}$$

Taking expectations and using the fact the X_i's are identically distributed we get,

$$ED(\hat{p}_{2n}||p) \leq ED(\hat{p}_n||p).$$

(b) The trick to this part is similar to part a) and involves rewriting $\hat{p}_n$ in terms of $\hat{p}_{n-1}$. We see that,

$$\hat{p}_n = \frac{1}{n}\sum_{i=0}^{n-1} I(X_i = x) + \frac{I(X_n = x)}{n}$$

or in general,

$$\hat{p}_n = \frac{1}{n}\sum_{i \neq j} I(X_i = x) + \frac{I(X_j = x)}{n},$$

where j ranges from 1 to n.
Summing over j we get,

$$n\hat{p}_n = \frac{n-1}{n}\sum_{j=1}^{n} \hat{p}^j_{n-1} + \hat{p}_n,$$

or,

$$\hat{p}_n = \frac{1}{n}\sum_{j=1}^{n} \hat{p}^j_{n-1}$$

where,

$$\sum_{j=1}^{n} \hat{p}_{n-1}^{j} = \frac{1}{n-1} \sum_{i \neq j} I(X_i = x).$$

Again using the convexity of $D(p||q)$ and the fact that the $D(\hat{p}_{n-1}^{j}||p)$ are identically distributed for all j and hence have the same expected value, we obtain the final result.

Chapter 3

(a) Since the $X_1, X_2, \ldots, X_n$ are i.i.d., so are $q(X_1), q(X_2), \ldots, q(X_n)$, and hence we can apply the strong law of large numbers to obtain

$$\begin{aligned}
\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n) &= \lim -\frac{1}{n}\sum \log q(X_i) && (3.7)\\
&= -E(\log q(X)) \text{ w.p. } 1 && (3.8)\\
&= -\sum p(x)\log q(x) && (3.9)\\
&= \sum p(x)\log\frac{p(x)}{q(x)} - \sum p(x)\log p(x) && (3.10)\\
&= D(\mathbf{p}||\mathbf{q}) + H(\mathbf{p}). && (3.11)
\end{aligned}$$

(b) Again, by the strong law of large numbers,

$$\begin{aligned}
\lim -\frac{1}{n}\log\frac{q(X_1, X_2, \ldots, X_n)}{p(X_1, X_2, \ldots, X_n)} &= \lim -\frac{1}{n}\sum \log\frac{q(X_i)}{p(X_i)} && (3.12)\\
&= -E(\log\frac{q(X)}{p(X)}) \text{ w.p. } 1 && (3.13)\\
&= -\sum p(x)\log\frac{q(x)}{p(x)} && (3.14)\\
&= \sum p(x)\log\frac{p(x)}{q(x)} && (3.15)\\
&= D(\mathbf{p}||\mathbf{q}). && (3.16)
\end{aligned}$$

1. *Markov's inequality and Chebyshev's inequality.*

(a) (Markov's inequality.) For any non-negative random variable X and any $\delta > 0$, show that

$$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta}. \tag{3.1}$$

Exhibit a random variable that achieves this inequality with equality.

(b) (Chebyshev's inequality.) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y - \mu)^2$, show that for any $\epsilon > 0$,

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \tag{3.2}$$

(c) (The weak law of large numbers.) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\overline{Z_n} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be the sample mean. Show that

$$\Pr\left\{\left|\overline{Z_n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{3.3}$$

Thus $\Pr\left\{\left|\overline{Z_n} - \mu\right| > \epsilon\right\} \to 0$ as $n \to \infty$. This is known as the weak law of large numbers.

Solution: *Markov's inequality and Chebyshev's inequality.*

(a) If X has distribution $F(x)$,

$$\begin{aligned} EX &= \int_0^\infty x \mathrm{d}F \\ &= \int_0^\delta x \mathrm{d}F + \int_\delta^\infty x \mathrm{d}F \end{aligned}$$

$$\begin{aligned} &\geq \int_{\delta}^{\infty} x \mathrm{d}F \\ &\geq \int_{\delta}^{\infty} \delta \mathrm{d}F \\ &= \delta \Pr\{X \geq \delta\}. \end{aligned}$$

Rearranging sides and dividing by δ we get,

$$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta}. \tag{3.4}$$

One student gave a proof based on conditional expectations. It goes like

$$\begin{aligned} EX &= E(X|X \leq \delta)\Pr\{X \geq \delta\} + E(X|X < \delta)\Pr\{X < \delta\} \\ &\geq E(X|X \leq \delta)\Pr\{X \geq \delta\} \\ &\geq \delta \Pr\{X \geq \delta\}, \end{aligned}$$

which leads to (3.4) as well.

Given δ, the distribution achieving

$$\Pr\{X \geq \delta\} = \frac{EX}{\delta},$$

is

$$X = \begin{cases} \delta & \text{with probability } \frac{\mu}{\delta} \\ 0 & \text{with probability } 1 - \frac{\mu}{\delta}, \end{cases}$$

where $\mu \leq \delta$.

(b) Letting $X = (Y - \mu)^2$ in Markov's inequality,

$$\begin{aligned} \Pr\{(Y-\mu)^2 > \epsilon^2\} &\leq \Pr\{(Y-\mu)^2 \geq \epsilon^2\} \\ &\leq \frac{E(Y-\mu)^2}{\epsilon^2} \\ &= \frac{\sigma^2}{\epsilon^2}, \end{aligned}$$

and noticing that $\Pr\{(Y-\mu)^2 > \epsilon^2\} = \Pr\{|Y-\mu| > \epsilon\}$, we get,

$$\Pr\{|Y-\mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

(c) Letting Y in Chebyshev's inequality from part (b) equal $\bar{Z}_n$, and noticing that $E\bar{Z}_n = \mu$ and $\mathrm{Var}(\bar{Z}_n) = \frac{\sigma^2}{n}$ (ie. $\bar{Z}_n$ is the sum of n iid r.v.'s, $\frac{Z_i}{n}$, each with variance $\frac{\sigma^2}{n^2}$), we have,

$$\Pr\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

3. *The AEP and source coding.* A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

   (a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

   (b) Calculate the probability of observing a source sequence for which no codeword has been assigned.

   (c) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (b).

Solution: *THe AEP and source coding.*

(a) The number of 100-bit binary sequences with *three or fewer* ones is

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751 .$$

The required codeword length is $\lceil \log_2 166751 \rceil = 18$. (Note that $H(0.005) = 0.0454$, so 18 is quite a bit larger than the 4.5 bits of entropy.)

(b) The probability that a 100-bit sequence has three or fewer ones is

$$\sum_{i=0}^{3} \binom{100}{i} (0.005)^i (0.995)^{100-i} = 0.60577 + 0.30441 + 0.7572 + 0.01243 = 0.99833$$

Thus the probability that the sequence that is generated cannot be encoded is $1 - 0.99833 = 0.00167$.

(c) In the case of a random variable S_n that is the sum of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, Chebyshev's inequality states that

$$\Pr(|S_n - n\mu| \geq \epsilon) \leq \frac{n\sigma^2}{\epsilon^2},$$

where μ and σ^2 are the mean and variance of X_i. (Therefore $n\mu$ and $n\sigma^2$ are the mean and variance of S_n.) In this problem, $n = 100$, $\mu = 0.005$, and $\sigma^2 = (0.005)(0.995)$. Note that $S_{100} \geq 4$ if and only if $|S_{100} - 100(0.005)| \geq 3.5$, so we should choose $\epsilon = 3.5$. Then

$$\Pr(S_{100} \geq 4) \leq \frac{100(0.005)(0.995)}{(3.5)^2} \approx 0.04061 .$$

This bound is much larger than the actual probability 0.00167.

5. *AEP.* Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

   (a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

   (b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n} \log \frac{q(X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favouring q are exponentially small when p is true.

Solution: (*AEP*).

Chapter 4

for $i = 1,3,7,9$, $H(X_2|X_1 = i) = \log 5$ for $i = 2,4,6,8$ and $H(X_2|X_1 = i) = \log 8$ bits for $i = 5$. Therefore, we can calculate the entropy rate of the king as

$$\begin{aligned} \mathcal{H} &= \sum_{i=1}^{9} \mu_i H(X_2|X_1 = i) && (4.65) \\ &= 0.3 \log 3 + 0.5 \log 5 + 0.2 \log 8 && (4.66) \\ &= 2.24 \text{ bits}. && (4.67) \end{aligned}$$

1. *Doubly stochastic matrices.* An $n \times n$ matrix $P = [P_{ij}]$ is said to be *doubly stochastic* if $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$ for all i and $\sum_i P_{ij} = 1$ for all j. An $n \times n$ matrix P is said to be a *permutation* matrix if it is doubly stochastic and there is precisely one $P_{ij} = 1$ in each row and each column.

   It can be shown that every doubly stochastic matrix can be written as the convex combination of permutation matrices.

   (a) Let $\mathbf{a}^t = (a_1, a_2, \ldots, a_n)$, $a_i \geq 0$, $\sum a_i = 1$, be a probability vector. Let $\mathbf{b} = \mathbf{a}P$, where P is doubly stochastic. Show that $\mathbf{b}$ is a probability vector and that $H(b_1, b_2, \ldots, b_n) \geq H(a_1, a_2, \ldots, a_n)$. Thus stochastic mixing increases entropy.

   (b) Show that a stationary distribution μ for a doubly stochastic matrix P is the uniform distribution.

   (c) Conversely, prove that if the uniform distribution is a stationary distribution for a Markov transition matrix P, then P is doubly stochastic.

   Solution: *Doubly Stochastic Matrices.*

   (a)

$$\begin{aligned} H(\mathbf{b}) - H(\mathbf{a}) &= -\sum_j b_j \log b_j + \sum_i a_i \log a_i && (4.1) \\ &= \sum_j \sum_i a_i P_{ij} \log(\sum_k a_k P_{kj}) + \sum_i a_i \log a_i && (4.2) \\ &= \sum_i \sum_j a_i P_{ij} \log \frac{a_i}{\sum_k a_k P_{kj}} && (4.3) \\ &\geq \left(\sum_{i,j} a_i P_{ij} \right) \log \frac{\sum_{i,j} a_i}{\sum_{i,j} b_j} && (4.4) \end{aligned}$$

$$= 1 \log \frac{m}{m} \tag{4.5}$$
$$= 0, \tag{4.6}$$

where the inequality follows from the log sum inequality.

(b) If the matrix is doubly stochastic, the substituting $\mu_i = \frac{1}{m}$, we can easily check that it satisfies $\mu = \mu P$.

(c) If the uniform is a stationary distribution, then

$$\frac{1}{m} = \mu_i = \sum_j \mu_j P_{ji} = \frac{1}{m} \sum_j P_{ji}, \tag{4.7}$$

or $\sum_j P_{ji} = 1$ or that the matrix is doubly stochastic.

2. *Time's arrow.* Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n).$$

In other words, the present has a conditional entropy given the past equal to the conditional entropy given the future.

This is true even though it is quite easy to concoct stationary random processes for which the flow into the future looks quite different from the flow into the past. That is to say, one can determine the direction of time by looking at a sample function of the process. Nonetheless, given the present state, the conditional uncertainty of the next symbol in the future is equal to the conditional uncertainty of the previous symbol in the past.

Solution: *Time's arrow.* By the chain rule for entropy,

$$H(X_0|X_{-1}, \ldots, X_{-n}) = H(X_0, X_{-1}, \ldots, X_{-n}) - H(X_{-1}, \ldots, X_{-n}) \tag{4.8}$$
$$= H(X_0, X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n) \tag{4.9}$$
$$= H(X_0|X_1, X_2, \ldots, X_n), \tag{4.10}$$

where (4.9) follows from stationarity.

4. *Monotonicity of entropy per element.* For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

(a)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.37}$$

(b)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n | X_{n-1}, \ldots, X_1). \tag{4.38}$$

Solution: *Monotonicity of entropy per element.*

(a) By the chain rule for entropy,

$$\begin{aligned} \frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^{n} H(X_i | X^{i-1})}{n} && (4.39) \\ &= \frac{H(X_n | X^{n-1}) + \sum_{i=1}^{n-1} H(X_i | X^{i-1})}{n} && (4.40) \\ &= \frac{H(X_n | X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. && (4.41) \end{aligned}$$

From stationarity it follows that for all $1 \leq i \leq n$,

$$H(X_n | X^{n-1}) \leq H(X_i | X^{i-1}),$$

which further implies, by averaging both sides, that,

$$\begin{aligned} H(X_n | X^{n-1}) &\leq \frac{\sum_{i=1}^{n-1} H(X_i | X^{i-1})}{n-1} && (4.42) \\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.43) \end{aligned}$$

Combining (4.41) and (4.43) yields,

$$\begin{aligned} \frac{H(X_1, X_2, \ldots, X_n)}{n} &\leq \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right] && (4.44) \\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.45) \end{aligned}$$

(b) By stationarity we have for all $1 \leq i \leq n$,

$$H(X_n | X^{n-1}) \leq H(X_i | X^{i-1}),$$

which implies that,

$$\begin{aligned} H(X_n | X^{n-1}) &= \frac{\sum_{i=1}^{n} H(X_n | X^{n-1})}{n} && (4.46) \\ &\leq \frac{\sum_{i=1}^{n} H(X_i | X^{i-1})}{n} && (4.47) \\ &= \frac{H(X_1, X_2, \ldots, X_n)}{n}. && (4.48) \end{aligned}$$

5. *Entropy rates of Markov chains.*

6. *Maximum entropy process.* A discrete memoryless source has alphabet $\{1, 2\}$ where the symbol 1 has duration 1 and the symbol 2 has duration 2. The probabilities of 1 and 2 are p_1 and p_2, respectively. Find the value of p_1 that maximizes the source entropy per unit time $H(X)/El_X$. What is the maximum value H?

Solution: *Maximum entropy process.* The entropy per symbol of the source is

$$H(p_1) = -p_1 \log p_1 - (1 - p_1) \log(1 - p_1)$$

and the average symbol duration (or time per symbol) is

$$T(p_1) = 1 \cdot p_1 + 2 \cdot p_2 = p_1 + 2(1 - p_1) = 2 - p_1 = 1 + p_2 .$$

Therefore the source entropy per unit time is

$$f(p_1) = \frac{H(p_1)}{T(p_1)} = \frac{-p_1 \log p_1 - (1 - p_1) \log(1 - p_1)}{2 - p_1} .$$

Since $f(0) = f(1) = 0$, the maximum value of $f(p_1)$ must occur for some point p_1 such that $0 < p_1 < 1$ and $\partial f / \partial p_1 = 0$.

$$\frac{\partial}{\partial p_1} \frac{H(p_1)}{T(p_1)} = \frac{T(\partial H/\partial p_1) - H(\partial T/\partial p_1)}{T^2}$$

After some calculus, we find that the numerator of the above expression (assuming natural logarithms) is

$$T(\partial H/\partial p_1) - H(\partial T/\partial p_1) = \ln(1 - p_1) - 2 \ln p_1 ,$$

which is zero when $1 - p_1 = p_1^2 = p_2$, that is, $p_1 = \frac{1}{2}(\sqrt{5} - 1) = 0.61803$, the reciprocal of the golden ratio, $\frac{1}{2}(\sqrt{5} + 1) = 1.61803$. The corresponding entropy per unit time is

$$\frac{H(p_1)}{T(p_1)} = \frac{-p_1 \log p_1 - p_1^2 \log p_1^2}{2 - p_1} = \frac{-(1 + p_1^2) \log p_1}{1 + p_1^2} = -\log p_1 = 0.69424 \text{ bits.}$$

Note that this result is the same as the maximum entropy rate for the Markov chain in problem #4(d) of homework #4. This is because a source in which every 1 must be followed by a 0 is equivalent to a source in which the symbol 1 has duration 2 and the symbol 0 has duration 1.

7. *Initial conditions.* Show, for a *stationary* Markov chain, that

$$H(X_0|X_n) \geq H(X_0|X_{n-1}) .$$

Thus initial conditions X_0 become more difficult to recover as the future X_n unfolds.

Solution: *Initial conditions.* For a Markov chain, by the data processing theorem, we have

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{4.49}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{4.50}$$

or $H(X_0|X_n)$ increases with n.

10. *The entropy rate of a dog looking for a bone.* A dog walks on the integers, possibly reversing direction at each step with probability $p = .1$. Let $X_0 = 0$. The first step is equally likely to be positive or negative. A typical walk might look like this:

$$(X_0, X_1, \ldots) = (0, -1, -2, -3, -4, -3, -2, -1, 0, 1, \ldots)$$

(a) Find $H(X_1, X_2, \ldots, X_n)$.

(b) Find the entropy rate of this browsing dog.

(c) What is the expected number of steps the dog takes before reversing direction?

Solution: *The entropy rate of a dog looking for a bone.*

(a) By the chain rule,

$$\begin{aligned} H(X_0, X_1, \ldots, X_n) &= \sum_{i=0}^{n} H(X_i|X^{i-1}) \\ &= H(X_0) + H(X_1|X_0) + \sum_{i=2}^{n} H(X_i|X_{i-1}, X_{i-2}), \end{aligned}$$

$\mathbf{X}(n) = [X_1(n)\ X_2(n)\ X_3(n)]^T$ satisfies the following recursion:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}, \tag{4.72}$$

with initial conditions $\mathbf{X}(1) = [1\ 1\ 0]^T$.

(c) Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.73}$$

Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \tag{4.74}$$

Using the eigenvalue decomposition of A for the case of distinct eigenvalues, we can write $A = U^{-1}\Lambda U$, where Λ is a matrix of eigenvalues. Then $A^{n-1} = U^{-1}\Lambda^{n-1}U$. Show that we can write

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.75}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ do not depend on n. For large n, this sum is doominated by the largest term. Therefore argue that for $i = 1, 2, 3$, we have

$$\frac{1}{n}\log X_i(n) \to \log \lambda, \tag{4.76}$$

where λ is the largest (positive) eigenvalue. Thus the number of sequences of length n grows as λ^n for large n. Calculate λ for the matrix A above.

(d) We will now take a different approach. Consider a Markov chain whose state diagram is the one given in part (a), but with arbirtary transition probabilities. Therefore the probability transition matrix of this Markov chain is

$$P = \begin{bmatrix} 0 & \alpha & 1 \\ 1 & 0 & 0 \\ 0 & 1-\alpha & 0 \end{bmatrix}. \tag{4.77}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]^T. \tag{4.78}$$

(e) Maximize the entropy rate of the Markov chain over choices of α. What is the maximum entropy rate of the chain?

(f) Compare the maximum entropy rate in part (e) with $\log \lambda$ in part (c). Why are the two answers the same?