Reading: Chapter 3 to 5 of Cover

1. Initial conditions [prob 7 on p75 of Cover]. Show, for a Markov chain, that

\[ H(X_0 | X_n) \geq H(X_0 | X_{n-1}) \]

Thus initial condition \( X_0 \) becomes more difficult to recover as the future \( X_n \) unfolds

**Solution:** Your Solution Here.

2. AEP [modified prob 5 on p58 of Cover]. 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 \( -\frac{1}{n} \log p(X_1, \ldots, X_n) \to H(X) \) in probability. Let \( q(x_1, \ldots, x_n) = \prod_{i=1}^{n} q(x_i) \), where \( q \) is another probability mass function on \( \{1, 2, \ldots, m\} \).

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

**Solution:** Your Solution Here.

(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 favoring \( q \) are exponentially small when \( p \) is true. Give and interpret the scenario when the log likelihood ratio is infinite.

**Solution:** Your Solution Here.

3. Random box size [prob 6 on p58 of Cover]. An \( n \)-dimensional rectangular box with sides \( X_1, X_2, X_3, \ldots, X_n \) is to be constructed. The volume is \( V_n = \prod_{i=1}^{n} X_i \). The edge length \( l \) of an \( n \)-cube with the same volume of the random box is \( l = V_n^{\frac{1}{n}} \). Let \( X_1, X_2, \ldots \) be i.i.d. uniform random variables over the unit interval \([0,1]\). Find \( \lim_{n \to \infty} V_n^{\frac{1}{n}} \) and compare to \( (EV_n)^{\frac{1}{n}} \). Clearly, the expected edge length does not capture the idea of the volume of the box.

**Solution:** Your Solution Here.
4. The AEP and source coding [prob 3 on p57 in Cover]. Optional

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.

**Solution:** Your Solution Here.

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

**Solution:** Your Solution Here.

(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:** Your Solution Here.

5. Proof of Theorem 3.3.1 [modified prob 7 on p58 of Cover]. Let \( X_1, X_2, \ldots, X_n \) be i.i.d. \( \sim p(x) \). Let \( B^{(n)}_\delta \subseteq \mathcal{X}^n \) such that \( \Pr(B^{(n)}_\delta) > 1 - \delta \). \( \delta \in [0, 1) \).

(a) Given any two sets \( A, B \) such that \( \Pr(A) > 1 - \epsilon_1 \) and \( \Pr(B) > 1 - \epsilon_2 \), show that \( \Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2 \). Hence, \( \Pr(A^{(n)} \cap B^{(n)}_\delta) > 1 - \epsilon - \delta \).

**Solution:** Your Solution Here.

(b) Using the above result, show that

\[
(\forall \delta \in [0, 1)) (\forall \delta' > 0) (\exists N \in \mathbb{N}) (\forall n > N) \left( \frac{1}{n} \log |B^{(n)}_\delta| > H - \delta' \right)
\]

Hence,

\[
(\forall \delta \in [0, 1)) \left( \min \left\{|B^{(n)}_\delta| : B^{(n)}_\delta \subseteq \mathcal{X}^n, \Pr(B^{(n)}_\delta) > 1 - \delta \right\} \leq |A^{(n)}_\epsilon| \quad \text{as} \quad \epsilon \to 0 \right)
\]

**Solution:** Your Solution Here.

6. Entropy rates of Markov Chain [prob 5 on p74 of Cover]. Optional

(a) Find the entropy rate of the two-state Markov chain with transition matrix

\[
P = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{10} & 1 - p_{10} \end{bmatrix}
\]
Solution: Your Solution Here.

(b) What values of $p_{01}, p_{10}$ maximize the rate of part a?

Solution: Your Solution Here.

(c) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 - p & p \\ 1 & 0 \end{bmatrix}$$

Solution: Your Solution Here.

(d) Find the maximum value of the entropy rate of the Markov chain of part c. We expect that the maximizing value of $p$ should be less than $1/2$, since the 0 state permits more information to be generated than the 1 state.

Solution: Your Solution Here.

(e) Let $N(t)$ be the number of allowable state sequences of length $t$ for the Markov chain of part c. Find $N(t)$ and calculate

$$H_0 := \lim_{t \to \infty} \frac{1}{t} \log N(t)$$

[Hint: Find a linear recurrence that expresses $N(t)$ in terms of $N(t - 1)$ and $N(t - 2)$.] Why is $H_0$ an upper bound on the entropy rate of the Markov chain? Compare $H_0$ with the maximum entropy found in part d.

Solution: Your Solution Here.

(f) How could we optimally encode an arbitrary random source with the allowable state sequences assuming that AEP holds? What is the minimum and expected codeword length per source symbol in terms of the source entropy rate $H$ and the quantity $H_0$?

Solution: Your Solution Here.

7. The entropy rate of a dog looking for a bone [prob 4 on p75 of Cover]. A dog walks on the integers, possibly reversing direction at each step with probability $p = 0.1$. 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 $H(X_1^n)$ where $X_1^n := (X_1, X_2, \ldots, X_n)$. 

Solution: Your Solution Here.

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

Solution: Your Solution Here.

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

Solution: Your Solution Here.

8. General AEP. Consider the stochastic process \( \{X_i\}_{i \in \mathbb{Z}^+} \).

(a) *Independent sequence.* Suppose that \( X_i \)'s are independent (not necessarily identically distributed) with \( \text{Var}[\log p(X_i)] \) bounded above by some constant \( \alpha \).

i. Express the expectation \( H_n := E[-\frac{1}{n} \log p(X_{n1})] \) in terms of the entropies \( H(X_i) \).

Solution: Your Solution Here.

ii. Argue that AEP holds if the limit \( H_\infty := \lim_{n \to \infty} H_n \) exists. Construct an example where the limit \( H_\infty \) does not exist (i.e. specify the sequence \( \{H_n\} \) in terms of \( n \) such that the average does not converge).

Solution: Your Solution Here.

iii. Consider \( X_i \)'s to be deterministically 0 if \( i \in \bigcup_{k \in \mathbb{N}} \{2^k, \ldots, 2^{k+1} - 1\} \), and i.i.d. Bernoulli random variables with probability 1/2 otherwise. Give \( H_\infty \) if it exists. If not, briefly explains why.

Solution: Your Solution Here.

(b) **Difficult:** *Hidden Markov.* Let \( \{S_i\}_{i \in \mathbb{Z}^+} \) be the stationary 2-state Markov process with transition probabilities \( p_{01} = 2p_{10} = 0.5 \). If \( S_i = 0 \), \( X_i \)'s are independent Bernoulli distributed random variables with probability 1/2. Otherwise, \( X_i \) will be deterministically 0.5. Calculate the entropy rate of \( \{X_i\} \)

[Hint: apply chain rule on \( H(X_{n1}, S_{n1}) \) and see if any term is 0.]

Solution: Your Solution Here.

(c) Briefly explain whether the following statement is true/false.

i. The optimal compression scheme that minimizes the expected codeword length for a general stochastic process is not known if the process does not have AEP.

Solution: Your Solution Here.
ii. (See the historical notes on p59 of Cover.) The most general process for which AEP holds is continuous stationary ergodic process and the type of convergence is almost sure convergence.

**Solution:** Your Solution Here.

(d) **Difficult and Optional.** Random length. Let \( \{L_i\}_{i \in \mathbb{Z}^+} \) be a set of i.i.d. geometric random variables with probability 1/2. \( X_j \) is Bernoulli random variable with probability \( 1/L_k \) for \( j \in [1 + \sum_{i=1}^{k-1} L_i, \sum_{i=1}^{k} L_i] \) where \( k \in \mathbb{Z}^+ \). Briefly explain what is wrong with the following calculation of the entropy rate of \( \{X_i\} \) assuming that AEP holds for this process.

Consider \( L_1^m \). As \( m \to \infty \), there will be almost surely a fraction \( 2^{-l} \) of \( L_i \)'s taking values of \( l \). Then, there will be almost surely a fraction \( l2^{-l}P[l>0] \) of \( X_i \)'s being Bernoulli(1/\( L_k \)) random variables. Hence, we can obtain the entropy rate as follows,

\[
H(\mathcal{X}) = \frac{\sum_{l>0} l2^{-l}H[1/L_k]}{\sum_{l>0} l2^{-l}} = \frac{1}{2} \sum_{l>0} l2^{-l}H[1/L_k] \approx 0.447
\]

**Solution:** Your Solution Here.

9. **Information divergence** Let \( \{X_i\}_{i \in \mathbb{Z}^+} \) be a set of i.i.d. random variables with probability distribution \( Q \), which is unknown. We would like to compress \( \{X_i\} \) by defining the hypothesized typical set \( A(n) \) based on the estimate \( \tilde{P}(x_1^n) \) after observing the \( n \)-sequence \( \tilde{x}_1^n \).

(a) Suppose \( X_i \)'s are Bernoulli random variable, where the actual generating distribution \( Q := \text{Bernoulli}(q) \), the hypothesized distribution based on \( x_1^n \) is \( \tilde{P}(x_1^n) := \text{Bernoulli}(p := \sum_i \tilde{x}_i/n) \), and the typical set defined using \( \tilde{P}(x_1^n) \) is \( A(n) := \{x_1^n : \left| -\frac{1}{n} \log \tilde{P}(x_1^n) - H(p) \right| \leq \epsilon \} \)

i. Show that \( x_1^n \in A(n) \iff \left| \sum_{i=1}^n x_i - p \right| \leq \frac{\epsilon}{\log \frac{1}{q-p}} \)

**Solution:** Your Solution Here.

ii. If \( p \neq 1/2 \), show that as \( \epsilon \to 0 \),

\[
\Pr(A^{(n)}_\epsilon) = Q(A^{(n)}_\epsilon) \approx 2^{-D[p||q]n}
\]

Briefly explain whether this result holds for \( p = 1/2 \).

**Solution:** Your Solution Here.

(b) **Difficult and Optional.** Strong typicality. Consider a general distribution \( Q \) on a finite support set \( \mathcal{X} \). Let \( N(\alpha;x_1^n) \) be the number of \( \alpha \in \mathcal{X} \) in the sample sequence \( x_1^n \). We
define the distribution $P_{x^n_1} : \alpha \in \mathcal{X} \mapsto N(\alpha; x^n_1)$ to be the type (or empirical distribution) of $x^n_1$. We define $A^{(n)}_{\epsilon}$ to be the strong typical set based on $P_{x^n_1}$ as follows,

$$A^{(n)}_{\epsilon} := \left\{ x^n_1 : \forall \alpha \in \mathcal{X}, \left| P_{\tilde{x}^{n}_1}(\alpha) - P_{x^n_1}(\alpha) \right| < \epsilon \right\}$$

As a side note, the strong typical set has all the desired property of the (weak) typical set. With this stronger version, the following parts will generalize the interpretation of information divergence beyond the Bernoulli distribution, and avoid the problem raised in part ii.

i. Define $\mathcal{P}_n := \{ P : x^n_1 \in \mathcal{X}^n \land P = P_{x^n_1} \}$ to be the class of all types. Show that

$$|\mathcal{P}_n| = \left( n + |\mathcal{X}| - 1 \right) \leq (n + 1)|\mathcal{X}|-1$$

**Solution:** Your Solution Here.

ii. Define $\mathcal{T}_P := \{ x^n_1 \in \mathcal{X}^n : P_{x^n_1} = P \}$ to be the class of type $P \in \mathcal{P}_n$. Show that

$$P_{x^n_1}(\mathcal{T}_{P_{x^n_1}}) = \max_{P \in \mathcal{P}_n} P_{x^n_1}(\mathcal{T}_P) \geq \frac{1}{|\mathcal{P}_n|}$$

**Solution:** Your Solution Here.

iii. Show using the previous result that

$$|\mathcal{T}_{P_{x^n_1}}| = P_{x^n_1}(\mathcal{T}_{P_{x^n_1}}) 2^{nH[P_{x^n_1}]}
\geq 2^{nH[P_{x^n_1}]}$$

**Solution:** Your Solution Here.

iv. Finally, show that $Q(\mathcal{T}_{P_{x^n_1}}) \geq 2^{-nD[P_{x^n_1} \| Q]}$ and so $Pr(A^{(n)}_{\epsilon}) := Q(A^{(n)}_{\epsilon}) \geq 2^{-nD[P_{x^n_1} \| Q]}$ as $\epsilon \to 0$.

**Solution:** Your Solution Here.