1. For a stationary random process $X$, show that

(a) $\lim_{n \to \infty} \frac{1}{n} H(X^n)$ exists.

*Hint:* Show that $\frac{1}{n} H(X^n)$ is non-increasing in $n$.

**Solution:**

$$nH(X^{n+1}) - (n + 1)H(X^n) = nH(X_{n+1}|X^n) + nH(X^n) - (n + 1)H(X^n)$$

$$= \sum_{i=1}^{n} [H(X_{n+1}|X^n) - H(X_i|X^{i-1})]$$

$$= \sum_{i=1}^{n} [H(X_{n+1}|X^n) - H(X_{n+1}|X_{n+2-i})]$$

$$= (a) \leq 0,$$

where (a) is because of the stationarity of $X$ and (b) comes from the fact that conditioning does not increase entropy.

(b) $\lim_{n \to \infty} \frac{1}{n} H(X^n)$ is equal to $\lim_{k \to \infty} H(X_0|X_{-k}^{-1})$.

**Solution:**

First note that $H(X_0|X_{-k}^{-1})$ is non-increasing in $k$ thus $\lim_{k \to \infty} H(X_0|X_{-k}^{-1})$ exists.

$$\frac{1}{n} H(X^n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i|X^{i-1})$$

$$= \frac{1}{n} \sum_{i=1}^{n} H(X_0|X_{-1}^{i-1})$$

Convince yourself that if a sequence converges to a limit, then the running average of that sequence converges to the same limit, which completes the proof.

(c) $\lim_{k \to \infty} H(X_0|X_{-k}^{-1})$ is also equal to $H(X_0|X_{-\infty}^{-1})$ (only for those who have taken measure theoretic in probability).

**Solution:** Here is a handy theorem we are going to use (please refer to Theorem 5.21 from "Probability" by Leo Breiman for a proof.)

**Theorem** Let $Z, Y_1, Y_2, \ldots$, be random variables on $(\Omega, \mathcal{F}, P)$, such that $E|Z| \leq \infty$. Then

$$E(Z|Y_1^n) \xrightarrow{n \to \infty} E(Z|Y_1^\infty) \text{ w.p.1}$$

Recall that $P_{X_0|X_{-k}}(x) = P(X_0 = x|X_{-k}^{-1}) = E[1_{\{X_0=x\}}|X_{-k}^{-1}], \forall x \in \mathcal{X}$. Apply the theorem, we have

$$\lim_{k \to \infty} P_{X_0|X_{-k}}(x) = P_{X_0|X_{-\infty}}(x) \quad \text{w.p.1, } \forall x \in \mathcal{X}$$
Entropy $H(P)$ is a bounded continuous function in $P$. Use bounded convergence theorem to complete the proof.

2. Entropy is essentially a lower bound on expected length of any lossless code:

(a) Give an example of a lossless code that does not satisfy the Kraft inequality.

**Solution:**
Let $\mathcal{X} = \{1, 2, 3\}$, $c(1) = 0, c(2) = 1, c(3) = 00$.

(b) Generalize and show that, in fact, the “Kraft sum” associated with a lossless code may be arbitrarily large (for sufficiently large source alphabet).

**Solution:** Let $\mathcal{X} = \{1, 2, \ldots, 2^{n+1} - 2\}$. Consider a node in a complete binary tree. The path from the root to the node gives a unique binary representation of that node. Call the binary representation the codeword associated with that node. Now use all the codewords associated with the nodes except the root in a complete binary tree with height $n$. We will have $2$ codeword with length $2$, $4$ with length $2$, ..., $2^i$ with length $i$, ....

The “Kraft sum” is $\sum_{i=1}^{n} 2^{2^{i-1}} = n$. Let $n \to \infty$, we have shown that the “Kraft sum” can be arbitrarily large.

Our proof that the entropy lower bounds the expected length of a uniquely decodable code heavily relied on the Kraft inequality. Indeed:

(c) Give an example of a lossless code for a random variable $X$ whose expected code-length is less than $H(X)$.

**Solution:** Let $\mathcal{X} = \{1, 2, 3\}$, $c(1) = 0, c(2) = 1, c(3) = 00$. Assign a uniform distribution of $\mathcal{X}$.

$E(L(X)) = 4/3$
$H(x) = \log 3 > 1.58 > 4/3$

In the remainder of this problem we will prove that, nevertheless, the entropy lower bounds the expected length of a general lossless code up to a small term (dependent on the size of the alphabet of $X$). Assume with no loss of generality that $X$ takes values in $\{1, \ldots, N\}$ and that $p_1 \geq p_2 \geq \cdots \geq p_N > 0$, where $p_i = \Pr\{X = i\}$. Let $L^{\text{opt}}$ denote the minimum expected code length among all lossless codes.

(d) Letting $l_i$ denote the length of the codeword for $i$, show that there exists a lossless source $L^{\text{opt}}$ for which $l_i = \lceil \log(i + 2)/2 \rceil$. Thus

$$L^{\text{opt}} = \sum_{i=1}^{N} p_i \lceil \log(i + 2)/2 \rceil.$$  

(e) Prove that

$$H(X) - \sum_{i=1}^{N} p_i \log(i + 2)/2 \leq \log \left( \sum_{i=1}^{N} \frac{2}{i+2} \right),$$
(f) Use previous items to show that
\[ L^{\text{opt}} \geq H(X) - \log(2\ln(N + 2)). \] (3)


3. The conditional empirical entropy of order \( k \) associated with \( x^n_{-(k-1)} \) is defined as:
\[ H_k(x^n_{-(k-1)}) = H(U_{k+1}|U^k), \]
where \( H(U_{k+1}|U^k) \) is the conditional entropy of random variable \( U_{k+1} \) given \( U^k \) with joint distribution \( P_{U_{k+1}|U^k}(u_{k+1}) = \frac{1}{n}|\{1 \leq i \leq n : x_i^i = u_{k+1}\}|. \)

Show That
(a) \( \min_{Q \in M_k} \frac{1}{n} \log \frac{1}{Q(x^n|x^0_{-(k-1)})} = H_k(x^n_{-(k-1)}) \)

Hint: The proof follows closely the way to prove \( \min_{Q \in M_0} \frac{1}{n} \log \frac{1}{Q(x^n)} = H_0(x^n). \)

Solution: For any \( Q \in M_k \), \( Q(x^n|x^0_{-(k+1)}) = \prod_{i=1}^{n} Q(x_i|x^i_{i-k}). \)

\[
\frac{1}{n} \log \frac{1}{Q(x^n|x^0_{-(k+1)})} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{Q(x_i|x^i_{i-k})} = \sum_{u_{k+1} \in \mathcal{X}^{k+1}} P_{U_{k+1}}(u_{k+1}) \log \frac{1}{Q(u_{k+1}|u^k)} \]
\[
= \sum_{u_k \in \mathcal{X}^k} P_{U_k}(u_k) \sum_{u_{k+1} \in \mathcal{X}} P_{U_{k+1}|U_k}(u_{k+1}) \log \left( \frac{P_{U_{k+1}|U_k}(u_{k+1})}{Q(u_{k+1}|u^k)} \cdot \frac{1}{P_{U_{k+1}}(u_{k+1})} \right) \]
\[
= \sum_{u_k \in \mathcal{X}^k} P_{U_k}(u_k) D(P_{U_{k+1}|U_k} || Q(\cdot|u^k)) + H(U_{k+1}|U^k) \geq H(U_{k+1}|U^k) = H_k(x^n_{-(k+1)}), \]
which completes the proof by the arbitrariness of \( Q \).

(b) For any stationary process \( X \), \( \mathbb{E} H_k(X^n_{-(k-1)}) \leq H(X_0|X_{-(k-1)}). \)

Hint: Jensen’s inequality.
4. We denote $N_{\text{LZ}}$ as the number of phrases in the LZ parsing of an individual sequence $x^n$.
Furthermore, denote $c_{l,u^k}$ as the number of phrases with length $l$ and left context $u^k$, i.e.,
\[ c_{l,u^k} = |\{ 1 \leq i \leq N_{\text{LZ}} : |y_i| = l, x_{v_{i+1}^n} = u^k \}|. \]

Refer to Lecture 8 for Lempel-Ziv compression.

(a) Let $L$ be a nonnegative integer-valued random variable with $\mathbb{E}L \leq \mu$, then
\[ H(L) \leq (\mu + 1) \log(\mu + 1) - \mu \log \mu \]

**Hint:** Prove that equality is attained when $L$ has a geometric distribution.

**Solution:** Let $G$ be a geometrical distributed random variable with pmf $Q(x) = p(1 - p)^x$ for $x = 0, 1, 2, \ldots$ and $p = \frac{1}{\mu + 1}$. Note that $\mathbb{E}G = \mu$. Consider
\[ 0 \leq D(P_L || Q) = \sum_{x=0}^{\infty} P_L(x) \log \frac{P_L(x)}{Q(x)} \]
\[ = -H(L) + \sum_{x=0}^{\infty} P_L(x) (\log p - x \log(1 - p)) \]
\[ \leq -H(L) - \log p - \mu \log(1 - p) \]
\[ = -H(L) + \log (\mu + 1) + \mu \log (\mu + 1) - \mu \log \mu \]
\[ = -H(L) + (\mu + 1) \log (\mu + 1) - \mu \log \mu \]

Thus, $H(L) \leq (\mu + 1) \log (\mu + 1) - \mu \log \mu$ and the equality holds iff $P_L(x) = Q(x)$ for all $x = 0, 1, 2, \ldots$, that is, $L$ has geometric distribution.
(b) Show that 

\[ N_{\text{LZ}} \leq \text{const} \frac{n}{\log n}, \]

where const is a constant independent of \( n \) and \( x^n \).

**Hint:** the lengths of the phrases are growing so one cannot pack too many of them in a sequence of length \( n \).

**Solution:** There are at most \( 2^l \) phrases of length \( l \). Let \( g(k) = \sum_{l=1}^{k} l \cdot 2^l = (k - 1)2^{k+1} + 2 \) and \( m(k) = \sum_{l=1}^{k} 2^l = 2^{k+1} - 2 \). Then \( g(k) \) is the minimum sequence length with \( m(k) \) distinct phrases. For a given \( n \), find a \( k \) such that \( g(k-1) < n \leq g(k) \). Then necessarily, \( N_{\text{LZ}} \leq m(k) \).

\[
N_{\text{LZ}} \frac{\log n}{n} \leq m(k) \frac{\log g(k)}{g(k-1)} \leq \frac{2^{k+1} \cdot 2(k+1)}{(k+1)2^{k-2}} = 16.
\]

In Cover&Thomas, a more accurate bound is derived

\[ N_{\text{LZ}} \leq \frac{n}{(1 - \epsilon_n) \log n}, \]

where \( \epsilon \to 0 \) as \( n \to \infty \).

(c) Show that

\[
\frac{N_{\text{LZ}}}{n} \sum_{l,u^k} c_{l,u^k} N_{\text{LZ}} \log N_{\text{LZ}} c_{l,u^k} \leq \epsilon^{(k)}_n,
\]

where \( \epsilon^{(k)}_n \) is independent of \( x^n \) and \( \lim_{n \to \infty} \epsilon^{(k)}_n = 0 \).

**Hint:** \( \sum_{l,u^k} c_{l,u^k} \log N_{\text{LZ}} c_{l,u^k} = H(L,U^k) \leq H(L) + H(U^k) \leq H(L) + k \log |\mathcal{X}| \), and now use the previous two parts.

**Solution:** Note that \( \sum_{l,u^k} c_{l,u^k}/N_{\text{LZ}} = 1 \). Let \((L,U^k)\) be a random vector with pmf. \( P_{L,U^k}(l,u^k) = c_{l,u^k}/N_{\text{LZ}} \). Then \( E L = \sum_{l,u^k} l \cdot c_{l,u^k}/N_{\text{LZ}} = n/N_{\text{LZ}} \)

\[
\sum_{l,u^k} c_{l,u^k} \frac{N_{\text{LZ}}}{c_{l,u^k}} = H(L,U^k) \leq H(L) + H(U^k) \leq H(L) + k \log |\mathcal{X}| = (\mu + 1) \log(\mu + 1) - \mu \log \mu + k \log |\mathcal{X}|,
\]

where \( \mu = \frac{n}{N_{\text{LZ}}} \).

\[
\frac{N_{\text{LZ}}}{n} \sum_{l,u^k} c_{l,u^k} \frac{N_{\text{LZ}}}{c_{l,u^k}} \leq \frac{1}{\mu} ((\mu + 1) \log(\mu + 1) - \mu \log \mu) + \frac{1}{\mu} k \log |\mathcal{X}| = \log \frac{\mu + 1}{\mu} + \frac{1}{\mu} \log(\mu + 1) + \frac{1}{\mu} k \log |\mathcal{X}| \triangleq \epsilon_n.
\]
From (b), we know that $\mu \to \infty$ as $n \to \infty$, which gives $\epsilon_n \to 0$ as $n \to \infty$. 