1. **Vin’s Idea (30 points)**

Vinith is very excited about a new lossless compression idea, and he claims it can beat entropy. Albert is very skeptical, as Vinith got a pretty low grade when he took EE376A. Albert, however, didn’t do too well in EE376A either, so now he needs your help to analyze Vinith’s scheme.

Vinith: “Suppose $X_1, X_2, \ldots$ is an i.i.d. Bernoulli-1/2 sequence. We can break up this sequence into its pattern of ‘repeats’. For instance, 0001100001... begins with repeats (also known as ‘run-lengths’) ‘000’, ‘11’, and ‘0000’. If we let $L_i$ be the length of the $i$th repeat, we can represent the sequence by $(X_1, L_1, L_2, \ldots)$. For example,

- 1010... would be represented by (1, 1, 1, 1, ...)
- 1110011110... by (1, 3, 2, 5, ...) and
- 00101111110... by (0, 2, 1, 1, 7, ...).

In particular, I suggest we describe the sequence $X_1, X_2, \ldots, X_{\sum_{i=1}^{10} L_i}$ by describing $(X_1, L_1, L_2, \ldots, L_{10})$, which I’m sure would be a heavily compressed representation!!”

(a) (5 points) What is the entropy of the first repeat length $H(L_1)$?

(b) (5 points) Describe an optimal prefix code for $L_1$. What is its expected code-length?

(c) (5 points) What is $H(X_1, L_1, \ldots, L_{10})$?

(d) (5 points) Describe an optimal uniquely decodable code for $(X_1, L_1, \ldots, L_{10})$. What is its expected code-length? Call it “Vinith’s code”.

(e) (5 points) What is the expected number of source symbols $E[\sum_{i=1}^{10} L_i]$ that Vinith’s code encodes?

(f) (5 points) Comment, based on your answers to the previous two parts, on whether Vinith’s code is “beating entropy” on average.

**Solution:**

(a) At each time $i$, the current repeat will end with probability 1/2. Therefore, each repeat length $L_i$ is a geometric random variable with parameter 1/2. Therefore,

$$H(L_1) = \sum_{j=1}^{\infty} 2^{-j} \log(2^j) = \sum_{j=1}^{\infty} j 2^{-j} = 2.$$  

(b) An optimum prefix code for $L_1$ is
\[ L_1 \quad \text{Codeword} \]
\[
1 \quad 1 \\
2 \quad 01 \\
3 \quad 001 \\
\vdots \quad \vdots
\]
where its expected code-length is
\[ \sum_{i=1}^{\infty} \frac{1}{2^i} = 2. \]

(c) Since the process is memoryless, the repeat lengths \( L_i \) are independent and identically distributed. Therefore \( H(L_1, \ldots, L_{10}) = 10H(L_1) = 20 \). The first symbol \( X_1 \) is independent of the repeat lengths \((L_1, \ldots, L_{10})\), so the joint entropy is given by
\[
H(X_1, L_1, \ldots, L_{10}) = H(X_1) + 10H(L_1) = 21.
\]

(d) We have to use one bit to describe \( X_1 \). Since \( L_1, L_2, \ldots, L_{10} \) are i.i.d. geometric random variables with parameter 1/2, we can use the prefix code for \( L_1 \) (which we did in (b)) repeatedly. Since the expected code-length of an optimal prefix code for \( L_1 \) is 2, the expected code-length will be
\[ 1 + 2 \times 10 = 21. \]

(e) The expected value of \( L_i \) is that of a geometric random variable with parameter 1/2:
\[ E[L_i] = 2. \]
Using the linearity of expectation,
\[
E[\sum_{i=1}^{10} L_i] = \sum_{i=1}^{10} E[L_i] = 20.
\]

(f) Vinith is encoding 20 source bits using an average of 21-bit-long binary codewords. Thus, on average he is expending more than one bit of description per source bit, and not ‘beating the entropy’.

2. Non-i.i.d. Source (30 points)
Consider a second-order binary Markov process \( \{X_i\}_{i \geq 1} \) characterized as follows:

- \( P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = \frac{1}{5} \) and 
  \( P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 0) = \frac{1}{3} \).
- For \( n \geq 3 \),
  - If \( X_{n-1} = X_{n-2} \), then \( X_n = 1 - X_{n-1} \).
  - If \( X_{n-1} \neq X_{n-2} \), then \( X_n \) is drawn as a fair coin flip, independent of \( \{X_i\}_{i=1}^{n-1} \).

(a) (6 points) Find an optimal prefix code for the pair \((X_1, X_2)\), along with its expected code-length.
(b) (6 points) Show that the distribution of \((X_n, X_{n+1})\) is the same for all \(n \geq 1\) (and, hence, the process is stationary).

(c) (6 points) Find the “entropy rate” of the process

\[
\lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}
\]

[Hint: can justify and use the facts that

\[H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X^{i-1})\]

and

\[H(X_i | X^{i-1}) = H(X_i | X_{i-1}, X_{i-2}) = H(X_3 | X_2, X_1) \text{ for } i \geq 3\]

(d) (6 points) For fixed \(n \geq 2\), does there exist a uniquely decodable code for \((X_1, X_2, \ldots, X_n)\) whose expected code-length is \(H(X_1, X_2, \ldots, X_n)\)? If so, describe one. If not, explain why.

(e) (6 points) Describe a uniquely decodable code for \((X_1, X_2, \ldots, X_n)\) that attains the entropy rate. That is, a code with length function \(\ell_n\) such that

\[
\lim_{n \to \infty} \frac{E[\ell_n(X_1, X_2, \ldots, X_n)]}{n}
\]

is equal to the entropy rate from part (c).

Solution:

(a) The Huffman tree for \((X_1, X_2)\) is

<table>
<thead>
<tr>
<th>Codeword</th>
<th>(X_1, X_2)</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(0, 1)</td>
<td>[\frac{1}{3}]</td>
<td>[\frac{1}{3}]</td>
</tr>
<tr>
<td>10</td>
<td>(1, 0)</td>
<td>[\frac{1}{3}]</td>
<td>[\frac{1}{3}]</td>
</tr>
<tr>
<td>110</td>
<td>(0, 0)</td>
<td>[\frac{1}{6}]</td>
<td>[\frac{1}{3}]</td>
</tr>
<tr>
<td>111</td>
<td>(1, 1)</td>
<td>[\frac{1}{6}]</td>
<td></td>
</tr>
</tbody>
</table>

Its expected code-length is

\[
1 \times \frac{1}{3} + 2 \times \frac{1}{3} + 3 \times \frac{1}{6} + 3 \times \frac{1}{6} = 2.
\]

(b) We will show that \((X_n, X_{n+1})\) has the same distribution with \((X_1, X_2)\) using induction. Clearly, the statement is true for \(n = 1\). Suppose \((X_{k-1}, X_k)\) has the same distribution with \((X_1, X_2)\). Then,

\[
P(X_k = 0, X_{k+1} = 0) = P(X_{k-1} = 0, X_k = 0, X_{k+1} = 0) + P(X_{k-1} = 1, X_k = 0, X_{k+1} = 0)
\]

\[
= P(X_{k+1} = 0 | X_{k-1} = 1, X_k = 0) P(X_{k-1} = 1, X_k = 0)
\]

\[
= \frac{1 \cdot 1}{2 \cdot 3}
\]

\[
= \frac{1}{6}
\]

\[
P(X_k = 0, X_{k+1} = 1) = P(X_{k-1} = 0, X_k = 0, X_{k+1} = 1) + P(X_{k-1} = 1, X_k = 0, X_{k+1} = 1)
\]
\[ P(X_{k+1} = 1|X_{k-1} = 0, X_k = 0)P(X_{k-1} = 0, X_k = 0) + P(X_{k+1} = 1|X_{k-1} = 1, X_k = 0)P(X_{k-1} = 1, X_k = 0) \]
\[ = 1 \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3} \]

Similarly, it is easy to show that \( P(X_k = 1, X_{k+1} = 0) = \frac{1}{3} \) and \( P(X_k = 1, X_{k+1} = 1) = \frac{1}{6} \).

(c) For \( n \geq 3 \), \( H(X_1, X_2, \ldots, X_n) = (n-2)H(X_3|X_2, X_1) + H(X_1, X_2) \). Therefore,
\[
\lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = H(X_3|X_2, X_1).
\]

Note that the conditional entropy \( H(X_3|X_2, X_1) \) is
\[
H(X_3|X_2, X_1) = \frac{1}{6} \cdot H(X_3|X_2 = 0, X_1 = 0) + \frac{1}{6} \cdot H(X_3|X_2 = 1, X_1 = 1) + \frac{1}{3} \cdot H(X_3|X_2 = 1, X_1 = 0) + \frac{1}{3} \cdot H(X_3|X_2 = 0, X_1 = 1)
\]
\[ = \frac{1}{6} \cdot 0 + \frac{1}{6} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{2}{3}. \]

(d) For \( n \geq 2 \),
\[ P(x_1, x_2, \ldots, x_n) = P(x_1, x_2) \prod_{i=3}^{n} P(x_i|x_{i-1}, x_{i-2}) \]

where \( P(x_1, x_2) \) is either \( \frac{1}{3} \) or \( \frac{1}{6} \), and \( P(x_i|x_{i-1}, x_{i-2}) \) takes value from \( \{1, 0, \frac{1}{2}\} \). Therefore, it is not a diadic distribution, we can not achieve \( H(X_1, X_2, \ldots, X_n) \).

(e) Consider the following coding scheme.

- Use any code for \( X_1, X_2 \) (e.g. Huffman). This will be negligible in terms of average code-length.
- For \( n \geq 3 \), if \( X_{n-1} \neq X_{n-2} \) describe \( X_n \) using 1 bit. If \( X_{n-1} = X_{n-2} \), then send nothing since \( X_n \) will be deterministic.

Since \( P(X_{n-1} = X_{n-2}) = \frac{2}{3} \), the average code-length per symbol will be \( \frac{2}{3} \).

Note: we can use Vin’s code to achieve the entropy rate. Let \( L_i \) be a length of \( i \)-th repeat. Then, \( Z_i = L_i - 1 \) is i.i.d. Bernoulli-1/2 random process and \((X_1, Z_1, Z_2, \ldots)\) will be a compressed version of \((X_1, X_2, \ldots)\). For given compressed version \((X_1, Z_1, Z_2, \ldots, Z_m)\), the expected number of encoded source symbols is
\[
E[L_1 + L_2 + \cdots + L_m] = \frac{3}{2} m.
\]

Therefore, we can argue that
\[
\lim_{n \to \infty} \frac{E[l_n(X_1, X_2, \ldots, X_n)]}{n} = \frac{2}{3}.
\]
3. Entropy of a Sum and a Difference of I.I.D. Random Variables (40 points)

We will prove that if \((Y, Y')\) are i.i.d. discrete random variables then:

\[
H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y)).
\]

We will prove this inequality in the following steps.

(a) **Data Processing Inequality for Mutual Information (10 points)**

Let \(X_1, X_2\) be discrete random variables. Also, let \(Y_1 = F(X_1)\) and \(Y_2 = G(X_2)\) for some functions, \(F(\cdot), G(\cdot)\). Prove that:

\[
I(X_1; X_2) \geq I(Y_1; Y_2).
\]

(b) **Submodularity (10 points)**

Suppose that there exist functions \(F, G\) and \(R\) such that \(X_0 = F(X_1) = G(X_2)\) and \(X_{12} = R(X_1, X_2)\), where \(X_1, X_2\) are discrete random variables. Use the previous part to prove that

\[
H(X_{12}) + H(X_0) \leq H(X_1) + H(X_2).
\]

(c) **Ruzsa Triangle Inequality (10 points)**

Let \(X, Y, Z\) be independent discrete random variables. Use Part (b) to prove that:

i. \(H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y)\)

ii. \(H(X - Z) \leq H(X + Y) + H(Y + Z) - H(Y)\)

[Hint: Use Part(b) with \(X_1 = (X - Y, Y - Z), X_2 = (X, Z), X_{12} = (X, Y, Z)\) and \(X_0 = X - Z\).]

(d) **Sum and Difference of Entropy (10 points)**

Use the previous part to conclude that for i.i.d \((Y, Y')\) random variables

\[
H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y)).
\]

Solution:

(a)

\[
I(X_1; X_2) = H(X_1) - H(X_1 | X_2) \\
\overset{(i)}{=} H(X_1) - H(X_1 | X_2, Y_2) \\
\overset{(ii)}{\geq} H(X_1) - H(X_1 | Y_2) \\
= I(Y_2; X_1) \\
= H(Y_2) - H(Y_2 | X_1) \\
\overset{(iii)}{=} H(Y_2) - H(Y_2 | X_1, Y_1)
\]
\[
\begin{align*}
\overset{(iv)}{\geq} & \quad H(Y_2) - H(Y_2 | Y_1) \\
= & \quad I(Y_1; Y_2),
\end{align*}
\]

where

(i) follows from the fact that \( Y_2 = G(X_2) \).

(ii) follows from conditioning reduces entropy.

(iii) follows from the fact that \( Y_1 = F(X_1) \).

(iv) follows from conditioning reduces entropy.

(b) Clearly, \( H(X_{12}) \leq H(X_1, X_2) \), thus \( H(X_1) + H(X_2) - H(X_{12}) \geq H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2) \). Proof is completed by using (a) above.

(c) i. Let \( X_1 = (X - Y, Y - Z) \), \( X_2 = (X, Z) \), \( X_0 = X - Z \) and \( X_{12} = (X, Y, Z) \). Thus we can have for some functions, \( F, G, R \) \( X_0 = F(X_1) = G(X_2) \) and \( X_{12} = R(X_1, X_2) \). Using (b) above we have:

\[ H(X, Y, Z) + H(X - Z) \leq H(X - Y, Y - Z) + H(X, Z) \]

Rearranging and using independence and conditioning reduces entropy,

\[ H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y). \]

ii. Replace \( Y \) by \(-Y\) and noting that \( H(Y) = H(-Y) \) we have the result.

(d) Use (c)-ii. for i.i.d. \( X, Y, Z \) and using \( H(Y) = H(X) \) and \( H(X + Y) = H(Y + Z) \), we get,

\[ H(X - Z) + H(X) \leq 2H(X + Y) \]

Now replace \( X = Y' \) and \( Z = Y'' \) to get the bound.