**EE 276 - Information Theory**
**Midterm**
**February 6, 2026**

1. You have 2 hours to take the exam. There are a total of 5 questions and 100 points. Questions have different numbers of points as indicated before each sub-problem.
2. Write your SUID (8-digit number) in the box on the top right corner of each page.
3. Please write your answer in the designated box underneath each question. **Answers written outside the box will not be graded.**
4. If you need more space for your work, you may use pages 14-15. In the original problem space, **you must indicate that you have used this extra space in order for it to be graded.**
5. All answers should be justified, unless otherwise stated.
6. Even if you did not prove a problem/subproblem, you may use its result in later problems/subproblems.
7. The exam is closed book but you are allowed one double-sided sheet of handwritten notes. No other materials are allowed.
8. Calculators are not allowed.
9. Do not discuss the contents of the exam with anyone who has not yet taken it.

Good luck!

**Full Name:** _____

**Email:** _____

**SUID:** _____

In accordance with both the letter and spirit of the Stanford Honor Code, I have neither given nor received unpermitted aid on this examination.

**Signature:** _____

1. 23 total points **Conditions for Finiteness of Entropy**

   Suppose $X$ is a random variable taking values in $\mathbb{N} = \{1, 2, \ldots\}$ with PMF $P$. We will attempt to evaluate the conditions under which the entropy of such a random variable is finite.

   (a) 7 points Suppose $\mathbb{E}[X] < \infty$. Let $Q$ be the PMF of a geometric random variable with success probability $r$, i.e., $Q(x) = (1-r)^{x-1}r \quad \forall x \in \mathbb{N}$ and $r \in (0, 1)$. Express $D(P||Q)$ in terms of $H(X)$, $\mathbb{E}[X]$, and $r$.

**Solution:** Expanding the KL-Divergence, we have that

$$\begin{aligned} D(P||Q) &= \sum_{x\in\mathbb{N}} P(x) \log \frac{P(x)}{Q(x)} = \sum_{x\in\mathbb{N}} P(x) \log P(x) + \sum_{x\in\mathbb{N}} P(x) \log \frac{1}{Q(x)} \\ &= -H(X) + \sum_{x\in\mathbb{N}} P(x) \log \frac{1}{(1-r)^{x-1}r} \\ &= -H(X) + (\mathbb{E}[X] - 1) \log \frac{1}{1-r} + \log \frac{1}{r} \end{aligned}$$

(b) [8 points] Use the previous part to show that

$$H(X) \leq \mathbb{E}[X] \cdot h_2\left(\frac{1}{\mathbb{E}[X]}\right),$$

with equality when $X$ is distributed as a geometric random variable. Hence, conclude that the entropy is finite. Here, $h_2(\cdot)$ denotes the binary entropy function given by $h_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$.

**Solution:** Let $Q$ be the same distribution as in the previous part. Using the fact that $D(P||Q) \geq 0$ we have that

$$H(X) \leq (\mathbb{E}[X]-1)\log\left(\frac{1}{1-r}\right) + \log\left(\frac{1}{r}\right).$$

This bound holds for any value of $r \in (0,1)$. Therefore, we shall pick the value of $r$ that minimizes the RHS. Differentiating with respect to $r$ and equating to 0, we get

$$\frac{\mathbb{E}[X]-1}{1-r} - \frac{1}{r} = 0 \implies r = \frac{1}{\mathbb{E}[X]}.$$

Substituting this in our upper bound, we get

$$\begin{aligned} H(X) &\leq (\mathbb{E}[X]-1)\log\left(\frac{1}{1-\frac{1}{\mathbb{E}[X]}}\right) + \log\left(\frac{1}{\frac{1}{\mathbb{E}[X]}}\right) \\ &= \mathbb{E}[X] \cdot \left[-\left(1-\frac{1}{\mathbb{E}[X]}\right)\log\left(1-\frac{1}{\mathbb{E}[X]}\right) - \frac{1}{\mathbb{E}[X]}\log\left(\frac{1}{\mathbb{E}[X]}\right)\right] \\ &= \mathbb{E}[X] \cdot h_2\left(\frac{1}{\mathbb{E}[X]}\right). \end{aligned}$$

Since the expectation is finite and the binary entropy function is bounded by 1, the entropy is upper bounded by a finite quantity, and is therefore finite. The upper bound is achieved when $D(P||Q) = 0$. This means that $P = Q$ and $X$ is distributed as a geometric random variable with parameter $r = \frac{1}{\mathbb{E}[X]}$.

(c) [8 points] Now, suppose we do not have that $\mathbb{E}[X] < \infty$ and instead have that $\mathbb{E}[\log X] < \infty$. Show that this implies $H(X) < \infty$.
*Hint: Consider a distribution $Q$ such that $\log \frac{1}{Q(x)} \propto \log x$. You may use the fact that $\sum_{x=1}^{\infty} \frac{1}{x^\alpha}$ converges to a finite value for any $\alpha > 1$.*

**Solution:** For any $\alpha > 1$, let

$$\sum_{x=1}^{\infty} \frac{1}{x^\alpha} = c_\alpha.$$

We know that such a finite $c_\alpha$ exists because this sum is convergent for $\alpha > 1$. Then we take our distribution $Q$ to be such that $Q(x) = \frac{1}{c_\alpha x^\alpha}$ for $x \in \mathbb{N}$. We can expand the KL-Divergence between $P$ and $Q$ as

$$D(P||Q) = -H(X) + \mathbb{E}\left[\log \frac{1}{Q(X)}\right].$$

Using the non-negativity of KL-Divergence, we have that

$$\begin{aligned} H(X) &\leq \mathbb{E}\left[\log c_\alpha X^\alpha\right] \\ &= \log c_\alpha + \alpha \mathbb{E}[\log X]. \end{aligned}$$

Since $c_\alpha$, $\alpha$, and $\mathbb{E}[\log X]$ are finite, the entropy is upper bounded by a finite quantity, and hence is finite.

**Note:** In this question we have shown that $\mathbb{E}[\log X] < \infty \implies H(X) < \infty$. In Homework 3 Question 4(c), we show that under the added condition of the monotonicity of $P$, $H(X) < \infty \implies \mathbb{E}[\log X] < \infty$.

2. [16 total points] **The Data-Processing Inequality for KL Divergence**

Let $\mathcal{X}, \mathcal{Y}$ be finite sets, and let $P_X, Q_X$ be probability mass functions over $\mathcal{X}$. Given a conditional PMF $P_{Y|X}(y|x)$, consider the two joint PMFs over $\mathcal{X} \times \mathcal{Y}$:

$$P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x),$$
$$Q_{X,Y}(x,y) = Q_X(x)P_{Y|X}(y|x).$$

**(Note that the conditional distribution of $Y$ given $X$ is the same under both of these joint PMFs.)**

In this problem we will prove the data-processing inequality for KL divergence.

(a) [8 points] (Monotonicity) Prove that

$$D(P_{X,Y}||Q_{X,Y}) \geq D(P_Y||Q_Y).$$

*Hint: The chain rule for KL divergence is potentially helpful, although it is not necessary.*

**Solution:** Applying the chain rule for KL divergence, we have that

$$D(P_{X,Y}||Q_{X,Y}) = D(P_Y||Q_Y) + D(P_{X|Y}||Q_{X|Y}|P_Y) \geq D(P_Y||Q_Y)$$

by nonnegativity of KL divergence.

(b) [8 points] (DPI for KL divergence) Prove that

$$D(P_X||Q_X) = D(P_{X,Y}||Q_{X,Y}),$$

and conclude that

$$D(P_Y||Q_Y) \leq D(P_X||Q_X). \tag{1}$$

**Solution:** By chain rule for KL divergence, we note that

$$D(P_{X,Y}||Q_{X,Y}) = D(P_X||Q_X) + D(P_{Y|X}||P_{Y|X}|P_X) = D(P_Y||Q_Y)$$

where the conditional KL divergence term is 0 because the conditional distribution is the same. We then can derive the inequality by combining the equality from this part with the inequality in part (a).

3. [21 total points] **Typical Sets for two Random Variables**

Let $p_X(0) = \frac{1}{2}, p_X(1) = p_X(2) = p_X(3) = \frac{1}{6}$. Let $p_Y(0) = \frac{3}{4}, p_Y(1) = p_Y(2) = p_Y(3) = \frac{1}{12}$. Let $X_1, X_2, \ldots$ be i.i.d. distributed according to $p_X$. Let $Y_1, Y_2, \ldots$ be i.i.d. distributed according to $p_Y$. We make no assumption about the relationship between the two sequences. Let $A_n := \mathcal{A}_\epsilon^{(n)}(X)$ and $B_n := \mathcal{A}_\epsilon^{(n)}(Y)$ be the typical sets for $X, Y$, respectively.

(a) [6 points] For a fixed $\epsilon > 0$, evaluate

$$\lim_{n\to\infty} \Pr\{X^n \in A_n \text{ and } Y^n \in B_n\}$$

**Solution:** By the AEP $\lim_{n\to\infty} \Pr\{X^n \in A_n\} = \lim_{n\to\infty} \Pr\{Y^n \in B_n\} = 1$. By the union bound,

$$\begin{aligned}\lim_{n\to\infty} \Pr\{X^n \in A_n \text{ and } Y^n \in B_n\} &= 1 - \lim_{n\to\infty} \Pr\{X^n \notin A_n \text{ or } Y^n \notin B_n\} \\ &\geq 1 - \lim_{n\to\infty} (\Pr\{X^n \notin A_n\} + \Pr\{X^n \notin A_n\}) \\ &= 1\end{aligned}$$

and hence the limit is 1.

(b) [7 points] For $\epsilon = 1/20$, evaluate

$$\lim_{n\to\infty} \Pr\{X^n \in A_n \cap B_n \text{ or } Y^n \in A_n \cap B_n\}$$

**Solution:** This is implied by the more general result in part (c).
More simply, we observe that $H(X) = 1+\log(3)/2$ and $H(Y) = H_2(3/4)+\log(3)/4$, where $H_2$ is the binary entropy function. Via a very rough bound $H(X) - H(Y) > \epsilon = 1/20$. For example, $H(X) - H(Y) > \log(3)/4 > 1/4 > 1/20$. Then, expanding the definitions, $A_n \cap B_n = \varnothing$, and the limit is therefore 0.

(c) [8 points] Now suppose $p_Z$ is such that $p_Z \neq p_X$ but $H(Z) = H(X)$. Let $Z_1, Z_2, \ldots$ be i.i.d. distributed according to $p_Z$. Prove or give a counterexample to the assertion that, for $\epsilon > 0$ sufficiently small, we have:

$$\lim_{n\to\infty} \Pr\{Z^n \in \mathcal{A}_\epsilon^{(n)}(X)\} = 0$$

**Solution:** The statement is true. By assumption the $p_Z \neq p_X$, we have that the relative entropy $D(p_Z \| p_X) > 0$. We then have that

$$\mathbb{E}[-\log p_X(Z)] = H(Z) + D(p_Z \| p_X) > H(X).$$

Let $\epsilon_0 = \mathbb{E}[-\log p_X(Z)] - H(X)$. By the WLLN, we have that

$$\lim_{n\to\infty} \Pr\left\{-\frac{1}{n}\sum_{i=1}^{n} \log p_X(Z_i) < \mathbb{E}[-\log p_X(Z)] - \epsilon_0/2\right\} = 0$$

$$\implies \lim_{n\to\infty} \Pr\left\{-\frac{1}{n}\sum_{i=1}^{n} \log p_X(Z_i) < H(X) + \epsilon_0/2\right\} = 0$$

$$\implies \lim_{n\to\infty} \Pr\{Z^n \in \mathcal{A}_{\epsilon_0/2}^n\} = 0$$

4. [24 total points] **Coin Toss Experiment and Golomb Codes**

Alice has been instructed to record the outcomes ($H$ or $T$) of a coin toss experiment. Consider the coin toss experiment $X_1, X_2, X_3, ...$ where $X_i$ are i.i.d. Bern($p$), i.e., probability of a head is $p$, where $p > 1/2$.

(a) [8 points] Alice suggests that as the occurrence of $T$ is so rare, we should just record the number of tosses it takes for each $T$ to occur. To be precise, if $Y_k$ represents the number of trials until the $k^{\text{th}}$ $T$ occurred (inclusive), then Alice records the sequence (with $Y_0 = 0$):

$$Z_k = Y_k - Y_{k-1}, \ k \geq 1.$$

Compute the expectation and entropy of $Z_k$. Compute the ratio of the entropy to the expectation as well, and give an intuitive explanation for the expression.
*Hint: You may find the result of Question 1(b) useful, but not necessary.*

**Solution:** $Z_k$ has geometric distribution with success probability $1 - p$. Specifically $\mathbb{P}\{Z_k = j\} = p^{j-1}(1-p)$ for $j \in \mathbb{N}$. To see this, let us consider $Z_1$ first. For $Z_1 = j$, we should have the first $j-1$ tosses be $H$, while the $j^{\text{th}}$ toss should be a $T$. This results in $\mathbb{P}\{Z_1 = j\} = p^{j-1}(1-p)$. Due to $Z_k$ being i.i.d., we can generalize the result to $Z_k$ for $k \geq 1$.

We have that $\mathbb{E}[Z_k] = \frac{1}{1-p}$. From the result of Question 1, we have that

$$\begin{aligned} H(Z_k) &= \mathbb{E}[Z_k] \cdot h_2\left(\frac{1}{\mathbb{E}[Z_k]}\right) \\ &= \frac{h_2(1-p)}{1-p} \end{aligned}$$

The ratio is $h_2(1-p) = h_2(p) = H(X)$. Intuitively, we can explain this as follows: $Z_k$ represents the same amount of information as $X_{Y_{k-1}+1}, ..., X_{Y_k}$. As the $X_i$ are i.i.d., the information content of $Z_k$ is the same as the information content of $E[Z_k]$ i.i.d random variables $X_i$. Thus, $H(Z_k) = E[Z_k]H(X)$

(b) [6 points] Consider the following scheme for encoding $Z_k$, which is a specific case of Golomb Coding. We are showing the first 10 codewords.

| Z | Quotient | Remainder | Code |
|---|---|---|---|
| 1 | 0 | 1 | 1 01 |
| 2 | 0 | 2 | 1 10 |
| 3 | 0 | 3 | 1 11 |
| 4 | 1 | 0 | 0 1 00 |
| 5 | 1 | 1 | 0 1 01 |
| 6 | 1 | 2 | 0 1 10 |
| 7 | 1 | 3 | 0 1 11 |
| 8 | 2 | 0 | 00 1 00 |
| 9 | 2 | 1 | 00 1 01 |
| 10 | 2 | 2 | 00 1 10 |

Describe the encoding and decoding schemes for this code.

**Solution:** We compute the quotient and remainder of $Z$ with respect to 4. Therefore, we can write $Z = 4q + r$ where $q \geq 0$ and $0 \leq r \leq 3$. The encoding of $Z$ is $q$ zeros followed by a 1, followed by the two-bit representation of $r$. When we decode, we count the number of zeros until we reach the first 1, which gives us $q$. We then decode the two bits after the 1 to give us $r$, thus obtaining $Z = 4q + r$.

(c) [10 points] Compute the expected codelength of this code.

**Solution:** To calculate the expected codelength, observe that the length of codeword for $Z = 4q + r$ (where $0 \leq r \leq 3$), is $3 + q$. Thus, adding terms in groups of 4 and subtracting one term corresponding to $Z = 0$, we get,

**Solution:** (Continued)

$$\begin{aligned}
\bar{l} &= \sum_{q=0}^{\infty}(1-p)p^{4q-1}(1+p+p^2+p^3)(3+q) - 3\frac{1-p}{p} \\
&= (1-p)(1+p+p^2+p^3)\sum_{q=0}^{\infty}p^{4q-1}(3+q) - 3\frac{1-p}{p} \\
&= (1-p)(1+p+p^2+p^3)\left[\frac{3}{p(1-p^4)} + p^3\sum_{q=0}^{\infty}p^{4q-4}q\right] - 3\frac{1-p}{p} \\
&= (1-p)(1+p+p^2+p^3)\left[\frac{3}{p(1-p^4)} + \frac{p^3}{(1-p^4)^2}\right] - 3\frac{1-p}{p} \\
&= (1-p)(1+p+p^2+p^3)\frac{3-2p^4}{p(1-p^4)^2} - 3\frac{1-p}{p} \\
&= \frac{3-2p^4}{p(1-p^4)} - 3\frac{1-p}{p}
\end{aligned}$$

5. 16 total points **Entropy of the English Language**

In this problem, we study the entropy of the English language as Shannon did. In what follows, we model an infinitely long string of English text as a random process

$$X_1, X_2, X_3 \ldots$$

where $X_i$ takes values in $\mathcal{X} := \{a, \ldots, z\}, |\mathcal{X}| = 26$. Here, $X_i$ represents the $i^{\text{th}}$ letter in a long English string. Note that the $X_i$'s are *not independent*, but for simplicity, we'll assume that the process is *stationary*, meaning that

$$p(x_1, \ldots, x_n) = p(x_{1+k}, \ldots, x_{n+k})$$

for all $n, k \geq 0$ and symbols $x_j \in \mathcal{X}$.

For $n \geq 1$, let us define the *n-gram entropy* as

$$\begin{aligned} \mathcal{H}_1 &:= H(X_1) \\ \mathcal{H}_n &:= H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1), \quad \text{for} \quad n \geq 2, \end{aligned}$$

which can be thought of as the entropy of the next letter in position $n$ given the previous $n-1$ letters. A natural definition of the entropy of English is then

$$\mathcal{H}_{\text{Eng}} := \lim_{n \to \infty} \mathcal{H}_n.$$

(a) 6 points Show that $\mathcal{H}_n \geq \mathcal{H}_{n+1}$.

**Solution:** The process is stationary. So, we should have $H(X_n | X_{n-1} X_{n-2} \ldots X_1) = H(X_{n+1} | X_n X_{n-1} \ldots X_2) = \mathcal{H}_n$. Since conditioning reduces entropy, we must have $\mathcal{H}_{n+1} = H(X_{n+1} | X_n X_{n-1} \ldots X_1) \leq \mathcal{H}_n$.

(b) [5 points] Using results from class, or otherwise, argue that $\mathcal{H}_n$ converges, i.e., that $\mathcal{H}_{\text{Eng}}$ exists and is finite.

**Solution:** We learned in class that this limit exists as long as the process is stationary. This can also be deduced from the previous part: namely, we know that $\mathcal{H}_0 \geq \mathcal{H}_n$. Furthermore, we know that conditional entropy is non-negative, so $\mathcal{H}_n \geq 0$. Due to the monotone convergence theorem, since $\mathcal{H}_n$ is bounded and monotonically decreasing, the limit exists.

(c) [5 points] For the sake of simplicity, in what follows, we'll only consider approximating $\mathcal{H}_{\text{Eng}}$ using $n = 1$ and $n = 2$. When Shannon and his wife Mary estimated $\mathcal{H}_1$ and $\mathcal{H}_2$ from English text, they found that

$$\mathcal{H}_1 \approx 4.14 \text{ bits} \quad \text{and} \quad \mathcal{H}_2 \approx 3.56 \text{ bits}.$$

You showed in part (a) that $\mathcal{H}_2 \leq \mathcal{H}_1$. What does the fact that $\mathcal{H}_2$ is considerably smaller than $\mathcal{H}_1$ tell us about letter pairs in English?

**Solution:** $\mathcal{H}_2$ is strictly less than $\mathcal{H}_1$ because English is contextual, English letters depend on what comes before them, and therefore not independent. This means that conditioning on the letters we have seen in the past, will strictly reduce the entropy.

**Additional space**: Clearly state the problem(s) for which you are using this space.

**Additional space**: Clearly state the problem(s) for which you are using this space.