Problem 1: Moment Generating Function

In the class we had considered the logarithmic moment generating function
\[ \phi(s) := \ln E[\exp(sX)] = \ln \sum_x p(x) \exp(sx) \]
of a real-valued random variable \( X \) taking values on a finite set, and showed that
\[ \phi'(s) = E[Xs] \]
where \( X_s \) is a random variable taking the same values as \( X \) but with probabilities
\[ p_s(x) := p(x) \exp(sx) \exp(-\phi(s)) . \]

(a) Show that
\[ \phi''(s) = \text{Var}(X_s) := E[X_s^2] - E[X_s]^2 \]
and conclude that \( \phi''(s) \geq 0 \) and the inequality is strict except when \( X \) is deterministic.

(b) Let \( x_{\min} := \min\{x : p(x) > 0\} \) and \( x_{\max} := \max\{x : p(x) > 0\} \) be the smallest and largest values \( X \) takes. Show that
\[ \lim_{s \to -\infty} \phi'(s) = x_{\min}, \quad \text{and} \quad \lim_{s \to \infty} \phi'(s) = x_{\max}. \]

Problem 2: Divergence and \( L_1 \)

Suppose \( p \) and \( q \) are two probability mass functions on a finite set \( \mathcal{U} \). (I.e., for all \( u \in \mathcal{U} \), \( p(u) \geq 0 \) and \( \sum_{u \in \mathcal{U}} p(u) = 1 \); similarly for \( q \).)

(a) Show that the \( L_1 \) distance \( \|p - q\|_1 := \sum_{u \in \mathcal{U}} |p(u) - q(u)| \) between \( p \) and \( q \) satisfies
\[ \|p - q\|_1 = 2 \max_{S,S' \subseteq \mathcal{U}} p(S) - q(S) \]
with \( p(S) = \sum_{u \in S} p(u) \) (and similarly for \( q \)), and the maximum is taken over all subsets \( S \) of \( \mathcal{U} \).

For \( \alpha \) and \( \beta \) in \([0,1]\), define the function \( d_2(\alpha\|\beta) := \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} \). Note that \( d_2(\alpha\|\beta) \) is the divergence of the distribution \((\alpha,1-\alpha)\) from the distribution \((\beta,1-\beta)\).

(b) Show that the first and second derivatives of \( d_2 \) with respect to its first argument \( \alpha \) satisfy
\[ d_2''(\beta\|\beta) = 0 \quad \text{and} \quad d_2''(\alpha\|\beta) = \frac{\log e}{\alpha(1-\alpha)} \geq 4 \log e . \]
(c) By Taylor’s theorem conclude that
\[ d_2(\alpha \| \beta) \geq 2(\log e)(\alpha - \beta)^2. \]

(d) Show that for any \( S \subset U \)
\[ D(p\|q) \geq d_2(p(S)\|q(S)) \]
[Hint: use the data processing theorem for divergence.]

(e) Combine (a), (c) and (d) to conclude that
\[ D(p\|q) \geq \frac{\log e}{4} \| p - q \|_1^2. \]

(f) Show, by example, that \( D(p\|q) \) can be \(+\infty\) even when \( \| p - q \|_1 \) is arbitrarily small. [Hint: considering \( U = \{0,1\} \) is sufficient.] Consequently, there is no generally valid inequality that upper bounds \( D(p\|q) \) in terms of \( \| p - q \|_1 \).

**Problem 3: Other Divergences**

Suppose \( f \) is a convex function defined on \((0, \infty)\) with \( f(1) = 0 \). Define the \( f \)-divergence of a distribution \( p \) from a distribution \( q \) as
\[
D_f(p\|q) := \sum_u q(u)f(p(u)/q(u)).
\]
In the sum above we take \( f(0) := \lim_{t \to 0} f(t) \), \( 0f(0/0) := 0 \), and \( 0f(a/0) := \lim_{t \to 0} tf(a/t) = a \lim_{t \to 0} tf(1/t) \).

(a) Show that for any non-negative \( a_1, a_2, b_1, b_2 \) and with \( A = a_1 + a_2, B = b_1 + b_2 \),
\[
b_1f(a_1/b_1) + b_2f(a_2/b_2) \geq Bf(A/B);
\]
and that in general, for any non-negative \( a_1, \ldots, a_k, b_1, \ldots, b_k \), and \( A = \sum_i a_i, B = \sum_i b_i \), we have
\[
\sum_i b_if(a_i/b_i) \geq Bf(A/B).
\]
[Hint: since \( f \) is convex, for any \( \lambda \in [0,1] \) and any \( x_1, x_2 > 0 \)
\( \lambda f(x_1) + (1 - \lambda)f(x_2) \geq f(\lambda x_1 + (1 - \lambda)x_2) \); consider \( \lambda = b_1/A \).]

(b) Show that \( D_f(p\|q) \geq 0 \).

(c) Show that \( D_f \) satisfies the data processing inequality: for any transition probability kernel \( W(v|u) \) from \( U \) to \( V \), and any two distributions \( p \) and \( q \) on \( U \)
\[
D_f(p\|q) \geq D_f(\tilde{p}\|\tilde{q})
\]
where \( \tilde{p} \) and \( \tilde{q} \) are probability distributions on \( V \) defined via \( \tilde{p}(v) := \sum_u W(v|u)p(u) \), and \( \tilde{q}(v) := \sum_u W(v|u)q(u) \).

(d) Show that each of the following are \( f \)-divergences.

i. \( D(p\|q) := \sum_u p(u) \log(p(u)/q(u)) \). [Warning: \( \log \) is not the right choice for \( f \).]
ii. \( R(p\|q) := D(q\|p) \).
iii. \( 1 - \sum_u \sqrt{p(u)q(u)} \)
iv. \( \| p - q \|_1 \)
v. \( \sum_u (p(u) - q(u))^2/q(u) \)
Problem 4: Entropy and pairwise independence

Suppose $X, Y, Z$ are pairwise independent fair flips, i.e., $I(X; Y) = I(Y; Z) = I(Z; X) = 0$.

(a) What is $H(X, Y)$?
(b) Give a lower bound to the value of $H(X, Y, Z)$.
(c) Give an example that achieves this bound.

Problem 5: Generating fair coin flips from biased coins

Suppose $X_1, X_2, \ldots$ are the outcomes of independent flips of a biased coin. Let $\Pr(X_i = 1) = p$, $\Pr(X_i = 0) = 1 - p$, with $p$ unknown. By processing this sequence we would like to obtain a sequence $Z_1, Z_2, \ldots$ of fair coin flips.

Consider the following method: We process the $X$ sequence in successive pairs, $(X_1, X_2), (X_3, X_4), (X_5, X_6)$, mapping $(01)$ to $0$, $(10)$ to $1$, and the other outcomes $(00)$ and $(11)$ to the empty string. After processing $X_1, X_2, \ldots$, we will obtain either nothing, or a bit $Z_1$.

(a) Show that, if a bit is obtained, it is fair, i.e., $\Pr(Z_1 = 0) = \Pr(Z_1 = 1) = 1/2$.
(b) With $h_2(p) = -p \log p - (1 - p) \log (1 - p)$, prove the following chain of (in)equalities.

\[
nh_2(p) = H(X_1, \ldots, X_n) \\
\geq H(Z_1, \ldots, Z_K, K) \\
= H(K) + H(Z_1, \ldots, Z_K | K) \\
= H(K) + E[K] \\
\geq E[K].
\]

Consequently, on the average no more than $nh_2(p)$ fair bits can be obtained from $(X_1, \ldots, X_n)$.

(c) Find a good $f$ for $n = 4$.

Problem 6: Extremal characterization for Rényi entropy

Given $s \geq 0$, and a random variable $U$ taking values in $\mathcal{U}$, with probabilities $p(u)$, consider the distribution $p_s(u) = p(u) / s$ with $Z(s) = \sum_u p(u)^s$.

(a) Show that for any distribution $q$ on $\mathcal{U}$,

\[
(1 - s)H(q) - sD(q||p) = -D(q||p_s) + \log Z(s).
\]

(b) Given $s$ and $p$, conclude that the left hand side above is maximized by the choice by $q = p_s$ with the value $\log Z(s)$.
The quantity
\[ H_s(p) := \frac{1}{1-s} \log Z(s) = \frac{1}{1-s} \log \sum_u p(u)^s \]
is known as the Rényi entropy of order \( s \) of the random variable \( U \). When convenient, we will also write \( H_s(U) \) instead of \( H_s(p) \).

(c) Show that if \( U \) and \( V \) are independent random variables
\[ H_s(UV) := H_s(U) + H_s(V). \]

[Here \( UV \) denotes the pair formed by the two random variables — not their product. E.g., if \( U = \{0,1\} \) and \( V = \{a,b\} \), \( UV \) takes values in \( \{0a,0b,1a,1b\} \).]

Problem 7: Guessing and Rényi entropy

Suppose \( X \) is a random variable taking values \( K \) values \( \{a_1, \ldots, a_K\} \) with \( p_i = \Pr\{X = a_i\} \). We wish to guess \( X \) by asking a sequence of binary questions of the type ‘Is \( X = a_i \)?’ until we are answered ‘yes’. (Think of guessing a password).

A guessing strategy is an ordering of the \( K \) possible values of \( X \); we first ask if \( X \) is the first value; then if it is the second value, etc. Thus the strategy is described by a function \( G(x) \in \{1, \ldots, K\} \) that gives the position (first, second, ..., \( K \) th) of \( x \) in the ordering. I.e., when \( X = x \), we ask \( G(x) \) questions to guess the value of \( X \). Call \( G \) the guessing function of the strategy.

For the rest of the problem suppose \( p_1 \geq p_2 \geq \cdots \geq p_K \).

(a) Show that for any guessing function \( G \), the probability of asking fewer than \( i \) questions satisfies
\[ \Pr(G(X) \leq i) \leq \sum_{j=1}^i p_j \]
and equality holds for the guessing function \( G^* \) with \( G^*(a_i) = i \), \( i = 1, \ldots, K \); this is the strategy that first guesses the most probable value \( a_1 \), then the next most probable value \( a_2 \), etc.

(b) Show that for any increasing function \( f : \{1, \ldots, K\} \to \mathbb{R} \), \( E[f(G(X))] \) is minimized by choosing \( G = G^* \). [Hint: \( E[f(G(X))] = \sum_{i=1}^K f(i) \Pr(G = i) \). Write \( \Pr(G = i) = \Pr(G \leq i) - \Pr(G \leq i-1) \), to write the expectation in terms of \( \sum_i [(f(i) - f(i+1)] \Pr(G \leq i) \), and use (a).]

(c) For any \( i \) and \( s \geq 0 \) prove the inequalities
\[ i \leq \sum_{j=1}^i (p_j/p_i)^s \leq \sum_j (p_j/p_i)^s \]

(d) For any \( \rho \geq 0 \), show that
\[ E[|G^*(X)|^\rho] \leq \left( \sum_i p_i^{1-s\rho} \right) \left( \sum_j p_j^s \right)^\rho. \]
for any \( s \geq 0 \). [Hint: write \( E[|G^*(X)|^\rho] = \sum_i p_i i^\rho \), and use (c) to upper bound \( i^\rho \)]

(e) By a choosing \( s \) carefully, show that
\[ E[|G^*(X)|^\rho] \leq \left( \sum_i p_i^{1/(1+\rho)} \right)^{1+\rho} = \exp[\rho H_1/(1+\rho)(X)]. \]
(f) Suppose $U_1, \ldots, U_n$ are i.i.d., each with distribution $p$, and $X = (U_1, \ldots, U_n)$. (I.e., we are trying to guess a password that is made of $n$ independently chosen letters.) Show that

$$\frac{1}{n} \log E[G^\rho(U_1, \ldots, U_n)^\rho] \leq H_{1/(1+\rho)}(U_1)$$

[Hint: first observe that $H_\alpha(X) = nH_\alpha(U_1)$. In other words, the $\rho$-th moment of the number of guesses grows exponentially in $n$ with a rate upper bounded by in terms of the Rényi entropy of the letters.

It is possible a lower bound to $E[G^\rho(U_1, \ldots, U_n)^\rho]$ that establishes that the exponential upper bound we found here is asymptotically tight.

**Problem 8: Gaussian variance estimation**

Consider estimating the mean $\mu$ and variance $\sigma^2$ from $n$ independent samples $(X_1, \ldots, X_n)$ of a Gaussian with this mean and variance.

(a) Show that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of $\mu$.

(b) Show that

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

is a biased estimator of $\sigma^2$ whereas

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of $\sigma^2$.

(c) Show that $S_n^2$ has a lower mean squared error than $S_{n-1}^2$. Thus it is possible that a biased estimator may be better than an unbiased one.