Problem 1. Convexity of Negative Entropy (4 points)

The negative entropy function is given by \( f(p) = \sum_{i=1}^{n} p_i \log p_i \) with \( p \geq 0 \). Show that this function is convex in four different ways:

**Problem 1(a)** By verifying that \( f \) satisfies the definition of a convex function.

We need to show that \( \sum_i (\theta x_i + (1 - \theta) y_i) \log(\theta x_i + (1 - \theta) y_i) \leq \theta \sum_i x_i \log x_i + (1 - \theta) \sum_i y_i \log y_i \) for any two \( x, y > 0 \). This is shown to be true by showing that \( (\theta x_i + (1 - \theta) y_i) \log(\theta x_i + (1 - \theta) y_i) \leq x_i \log x_i + (1 - \theta) y_i \log y_i \) holds for any \( i \). Let \( a_1 = \theta x_i, a_2 = (1 - \theta) y_i \). Also let \( b_1 = \theta, b_2 = (1 - \theta) \). Then the result follows by applying the log-sum inequality (Note that the log-sum inequality itself follows \( D(p||q) \geq 0 \) which follows from the convexity of \( -\log x \) - Which is easy to see/show).

**Problem 1(b)** By checking the first order condition for convexity: \( f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \).

The R.H.S is given by

\[
\begin{align*}
    f(x) + \langle \nabla f(x), y - x \rangle &= \sum_i x_i \log x_i + \sum_i \log x_i (y_i - x_i) \\
    &= \sum_i (y_i + y_i \log x_i - x_i) \\
    &= \sum_i g(x_i, y_i)
\end{align*}
\]

Consider the function \( g(x, y) = y + y \log x - x \). By taking the first and second derivatives w.r.t \( x \) it is clear that, \( g \) attains its maximum at \( x = y \). Thus \( g(x, y) \geq g(y, y) = y \log y \) and hence the result follows.

**Problem 1(c)** By checking the second order condition for convexity: \( \nabla^2 f(x) \succeq 0 \).

Let \( g(x) = x \log x \). \( g'(x) = 1 + \log x \). \( g''(x) = 1/x > 0 \). Since \( f(x) = \sum_i g(x_i) \) it follows that the hessian is a diagonal with the \( i^{th} \) entry given by \( 1/x_i \). Since the diagonals are positive, it follows that \( \nabla^2 f(x) \succeq 0 \).
Problem 1(d) By using the fact that sum of convex functions is convex.

This is trivially true since \( f(x) = \sum_i g(x_i) \) and we have already seen that \( g \) is convex.

(You may refer to Lecture 0 for definitions and some examples. For more details, a good reference is Section 3.1, Chapter 3 of “Convex Optimization” by Stephen Boyd and Lieven Vandenberghe: http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf).

Problem 2. Minimum entropy (3 points) 2.3

What is the minimum value of \( H(p_1, \ldots, p_n) = H(p) \) as \( p \) ranges over the set of \( n \)-dimensional probability vectors? Find all \( p \)'s which achieve this minimum.

We wish to find all probability vectors \( p = (p_1, p_2, \ldots, p_n) \) which minimize

\[
H(p) = -\sum_i p_i \log p_i.
\]

Now \( -p_i \log p_i \geq 0 \), with equality iff \( p_i = 0 \) or 1. Hence the only possible probability vectors which minimize \( H(p) \) are those with \( p_i = 1 \) for some \( i \) and \( p_j = 0, j \neq i \). There are \( n \) such vectors, i.e., \((1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)\), and the minimum value of \( H(p) \) is 0.

Problem 3. Drawing with and without replacement (3 points) 2.8

An urn contains \( r \) red, \( w \) white, and \( b \) black balls. Which has higher entropy, drawing \( k \geq 2 \) balls from the urn with replacement or without replacement? Set it up and show why. (There is both a hard way and a relatively simple way to do this.)

Drawing with and without replacement. Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the \( i \)-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

- With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

\[
X_i = \begin{cases} 
\text{red} & \text{with prob.} \frac{r}{r+w+b} \\
\text{white} & \text{with prob.} \frac{w}{r+w+b} \\
\text{black} & \text{with prob.} \frac{b}{r+w+b}
\end{cases}
\]

and therefore

\[
H(X_i|X_{i-1}, \ldots, X_1) = H(X_i) = \log(r + w + b) - \frac{r}{r+w+b} \log r - \frac{w}{r+w+b} \log w - \frac{b}{r+w+b} \log b.
\]
• Without replacement. The unconditional probability of the i-th ball being red is still \( r/(r+w+b) \), etc. Thus the unconditional entropy \( H(X_i) \) is still the same as with replacement. The conditional entropy \( H(X_i|X_{i-1}, \ldots, X_1) \) is less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

**Problem 4. A measure of correlation (4 points) 2.11**

Let \( X_1 \) and \( X_2 \) be identically distributed, but not necessarily independent. Let

\[
\rho = 1 - \frac{H(X_2 | X_1)}{H(X_1)}.
\]

1. Show \( \rho = \frac{H(X_1; X_2)}{H(X_1)} \).
2. Show \( 0 \leq \rho \leq 1 \).
3. When is \( \rho = 0 \)?
4. When is \( \rho = 1 \)?

**Problem 5. Entropy of Indicators (7 points)**

Let \( Z \sim U[0, 1] \), i.e. \( Z \) is a continuous random variable with uniform distribution over the interval \([0, 1]\). Note that the probability density function (p.d.f) of \( Z \) is given by \( f_Z(z) = 1 \forall 0 \leq z \leq 1 \). Also let \( Z_n \) (for \( n \) - positive integer) be a discretization of \( Z \) with
probability mass function (p.m.f) as follows: \( P(Z_n = \frac{i}{n}) = \frac{1}{n+1} \) for all integers \( i \) in \([0, n]\). Now define an indicator random variable \( X_n \in \{0, 1\} \) w.r.t \( Z_n \) as follows:

\[
X_n = \begin{cases} 
0 & \text{with probability } p \text{ if } 0 \leq Z_n \leq 1/4 \\
1 & \text{with probability } q \text{ if } 1/4 < Z_n \leq 1
\end{cases}
\]

I.e. \( X_n \) probabilistically indicates a value in \( \{0, 1\} \) whenever \( Z_n \) falls in a certain range. For the first three subproblems below, assume that \( p = 1/4, q = 1/4 \) in (4).

**Problem 5(a)** Evaluate \( E[X_n] \).

(4) can be restated as the following conditional probabilities:

\[
P(X_n = 0|0 \leq Z_n \leq 1/4) = p, \quad P(X_n = 1|0 \leq Z_n \leq 1/4) = 1 - p
\]

\[
P(X_n = 0|1/4 < Z_n \leq 1) = 1 - q, \quad P(X_n = 1|0 < Z_n \leq 1) = q
\]

Also note that \( P(0 \leq Z_n \leq 1/4) = \frac{1}{n+1} \lfloor \frac{n}{4} \rfloor + 1 \).

By the law of total expectation,

\[
E[X_n] = E[E[X_n|Z_n]] = P(0 \leq Z_n \leq 1/4)E[X_n|0 \leq Z_n \leq 1/4] + P(1/4 < Z_n \leq 1)E[X_n|1/4 < Z_n \leq 1] = \frac{\lfloor \frac{n}{4} \rfloor + 1}{n+1} (1-p) + \frac{n-\lfloor \frac{n}{4} \rfloor}{n+1} q
\]

**Problem 5(b)** How likely is \( X_n = 0 \) an indicator of \( Z_n \) being greater than 1/4? What is the limit of this probability?

The problem can be rephrased as the conditional probability: \( P(Z_n > 1/4|X_n = 0) \). By Bayes rule,

\[
P(Z_n > 1/4|X_n = 0) = \frac{P(X_n=0|Z_n>1/4)P(Z_n>1/4)}{P(X_n=0)} = \frac{(1-q)^n - \frac{\lfloor n/4 \rfloor}{n+1}}{p \frac{\lfloor n/4 \rfloor + 1}{n+1} + (1-q)^n - \frac{\lfloor n/4 \rfloor}{n+1}}
\]

As \( n \to \infty \), \( P(Z_n > 1/4|X_n = 0) \to \frac{3(1-q)}{4} \frac{3q}{4} = \frac{3-3q}{p+3-q} \). For \( p = 1/4, q = 1/4 \) we have that

\[
P(Z_n > 1/4|X_n = 0) \to \frac{9}{10}.
\]

**Problem 5(c)** Evaluate \( H(X_n), H(X_n|Z_n) \). Let \( X = \lim X_n \). Evaluate \( H(X) \). Note that we have not yet defined mutual information for continuous random variables. Nevertheless, evaluate \( I(X; Z) \) as the limit of \( I(X_n; Z_n) \).

Note that \( P(X_n = 0) = p \frac{\lfloor n/4 \rfloor + 1}{n+1} + (1-q)^n - \frac{\lfloor n/4 \rfloor}{n+1} \). Thus \( H(X_n) = -P(X_n = 0) \log P(X_n = 0) - P(X_n = 1) \log P(X_n = 1) \). \( H(X_n|Z_n) = P(0 \leq Z_n \leq 1/4)H(X_n|0 \leq Z_n \leq 1/4) + P(1/4 < Z_n \leq 1)H(X_n|1/4 < Z_n \leq 1) \). \( H(X_n; Z_n) = H(X_n) - H(X_n|Z_n) \). Note that \( P(X = 0) = \frac{p}{4} + \frac{3(1-q)}{4} \). For
\[ p = q = 1/4, \text{ this evaluates to } P(X = 0) = 5/8. \text{ Thus } H(X) = -5/8 \log 5/8 - 3/8 \log 3/8 = H(3/8). \]

Also \( H(X|Z) = \lim H(X_n|Z_n) = \frac{1}{4} H(p) + \frac{3}{4} H(q). \) For \( p = q = 1/4, \ H(X|Z) = H(1/4). \) Therefore, \( I(X; Z) = \lim I(X_n; Z_n) = H(X) - H(X|Z) = H(3/8) - H(1/4) = 0.143. \)

**Problem 5(d)** Now let \( p = 1/4, q = 3/4. \) Repeat problem 5c). What can be said about the relationship between \( X \) and \( Z \) in these two cases?

With \( p = 1/4, q = 3/4, \ P(X = 0) = \frac{1}{4}. \text{ Thus, } H(X) = H(1/4). \text{ Also, } H(X|Z) = \frac{1}{4} H(1/4) + \frac{3}{4} H(3/4) = H(1/4). \text{ Therefore, } I(X; Z) = H(X) - H(X|Z) = 0. \text{ With } p = 1/4, q = 1/4, \ X \text{ and } Z \text{ share some mutual information of 0.143: Thus } X_n \text{ is not a bad indicator of } Z_n. \text{ With } p = 1/4, q = 3/4, \text{ no mutual information is shared. Thus in this case } X_n \text{ is a bad indicator of } Z_n \text{ since it behaves as if it is independent of } Z_n (\text{indicating zero information about } Z_n). \text{ The best case is when } p = 1, q = 1. \text{ In this case the mutual information equals } H(1/4) = 0.81 \text{ which is much higher than the previous two cases. Also in this case, } X_n \text{ indicates full information about the range of } Z_n.

**Problem 6. Bounding the natural log (4 points)**

Let \( f(x) = \ln x - 1 + \frac{1}{x}. \) Show that \( f(x) \geq 0 \) for all \( x > 0 \) in two ways as described below.

**Problem 6(a)** First show that \( f(1) = 0 \) and that \( f'(x) \) is increasing for \( x \geq 1. \) Finish the argument to show that \( f(x) \geq 0 \) for all \( x > 0. \) List all \( x \) for which \( f(x) = 0. \)

\[
f(1) = \ln 1 - 1 + 1 = 0. \quad f'(x) = \frac{1}{x} - \frac{1}{x^2}. \quad \text{For } x \geq 1, x^2 \geq x \text{ and hence } f'(x) \geq 0 \text{ for } x \geq 1. \quad \text{Similarly for } x \leq 1, x^2 \leq 1 \text{ and hence } f'(x) \leq 0 \text{ for } x \leq 1. \quad \text{Thus, clearly the minimum is attained at } x = 0. \quad \text{Therefore } f(x) \geq 0 \forall x > 0. \quad \text{Also, } f(x) = 0 \text{ only at } x = 1.
\]

**Problem 6(b)** Starting from the result that was shown in class: \( e^x \geq 1 + x \forall x, \) argue that \( f(x) \geq 0 \) for all \( x > 0. \)

\[
e^x \geq 1 + x \text{ implies that } e^{x-1} \geq x. \text{ Thus, for all } x > 0, x - 1 \geq \ln x. \text{ Let } x = 1/y. \text{ Then for all } y > 0, 1/y - 1 \geq -\ln y. \text{ I.e. } \ln y - 1 + \frac{1}{y} = f(y) \geq 0 \forall y > 0.
\]

**Problem 7. Average entropy (3 points) 2.24 a), b)**

Let \( H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \) be the binary entropy function.

1. Evaluate \( H(1/4) \) using the fact that \( \log_2 3 \approx 1.584. \) **Hint:** You may wish to consider an experiment with four equally likely outcomes, one of which is more interesting than the others.

2. Calculate the average entropy \( H(p) \) when the probability \( p \) is chosen uniformly in the range \( 0 \leq p \leq 1. \)
**Hint:** Let \( E[X|Y] \) be the expectation taken w.r.t random variable \( X \) given random variable \( Y \). Note that \( E[X|Y] \) is itself a random variable that is a function of \( Y \). You may use the fact that \( E[X] = E[E[X|Y]] \). Note the difference in the definitions of \( E[X|Y] \) with \( H(X|Y) \): one is a random function of \( Y \), the other is not.

---

**Average Entropy.**

1. We can generate two bits of information by picking one of four equally likely alternatives. This selection can be made in two steps. First we decide whether the first outcome occurs. Since this has probability \( \frac{1}{4} \), the information generated is \( H(1/4) \). If not the first outcome, then we select one of the three remaining outcomes; with probability \( \frac{3}{4} \), this produces \( \log_2 3 \) bits of information. Thus

\[
H(1/4) + (3/4) \log_2 3 = 2
\]

and so \( H(1/4) = 2 - (3/4) \log_2 3 = 2 - (.75)(1.585) = 0.811 \) bits.

2. If \( p \) is chosen uniformly in the range \( 0 \leq p \leq 1 \), then the average entropy (in nats) is

\[
-\int_0^1 p \ln p + (1-p) \ln(1-p)dp = -2\int_0^1 x \ln x dx = -2 \left( \frac{x^2}{2} \ln x + \frac{x^2}{4} \right) \bigg|_0^1 = \frac{1}{2}.
\]

Therefore the average entropy is \( \frac{1}{2} \log_2 e = 1/(2 \ln 2) = .721 \) bits.

---

**Problem 8. Mixing increases entropy (2 points) 2.28**

**Problem 8(a)**  Show that the entropy of the probability distribution,  
\((p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m)\), is less than the entropy of the distribution  
\((p_1, \ldots, p_i + p_j/2, \ldots, p_i + p_j/2, \ldots, p_m)\).

**Problem 8(b) Bonus (2 points)**  Show that in general any transfer of probability that makes the distribution more uniform increases the entropy.

---

**Mixing increases entropy.**

This problem depends on the convexity of the log function. Let

\[
\begin{align*}
P_1 &= (p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m) \\
P_2 &= (p_1, \ldots, p_i + p_j/2, \ldots, p_j + p_i/2, \ldots, p_m)
\end{align*}
\]

Then, by the log sum inequality,

\[
H(P_2) - H(P_1) = -2(p_i + p_j/2) \log(p_i + p_j/2) + p_i \log p_i + p_j \log p_j \\
= -(p_i + p_j) \log(p_i + p_j/2) + p_i \log p_i + p_j \log p_j \\
\geq 0.
\]
Thus,

\[ H(P_2) \geq H(P_1). \]