Instructions: This is an extra problem set, not for submission.

1) Invariance to bijection: Let $X$ and $Y$ be i.i.d. according to $\text{Unif}\{\{0, 1, 2, 3\}\}$.
   a) Prove that $H(X + 4Y) = H(X, Y)$.
   b) Calculate $H(X + 4Y)$.

2) KL-divergence computation: Compute the following $f$-divergences:
   a) $D_{KL}(\text{Exp}(\eta_1) || \text{Exp}(\eta_2))$, where $\text{Exp}(\eta)$ is the exponential distribution with parameter $\eta > 0$, i.e., the distribution whose PDF is $p(\eta)(x) = \eta e^{-\eta x} 1_{\{x \geq 0\}}$.
   b) $D_{KL}(\text{Bin}(n, p_1) || \text{Bin}(n, p_2))$, where $\text{Bin}(n, p)$ is the Binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$, i.e., the distribution whose PMF is $p(n, p)(k) = \binom{n}{k} p^k (1 - p)^{n-k}$, for $k \in \{0, 1, 2, \ldots, n\}$ (otherwise 0).
   c) $D_{KL}(\text{Geo}(p_1) || \text{Geo}(p_2))$, where $\text{Geo}(p)$ is the geometric distribution with parameter $p \in [0, 1]$, i.e., the distribution whose PMF is $p(p)(k) = (1 - p)^k p$, for $k \in \mathbb{N}$ (otherwise 0).

3) Shannon entropy and KL divergence
   Consider the following two distributions $P$ and $Q$ supported on $\{1, 2, \ldots, L + M\}$, with PMFs:
   $$ p(x) = \begin{cases} p_x, & x = 1, \ldots, L; \\ 0, & x = L + 1, \ldots, L + M \end{cases} $$
   $$ q(x) = \begin{cases} \alpha p_x, & x = 1, \ldots, L; \\ \frac{1-\alpha}{M}, & x = L + 1, \ldots, L + M \end{cases} $$
   where $0 < \alpha < 1$.
   a) Compute $D_{KL}(P || Q)$
   b) Express $H(Q)$ in terms of $H(P)$, $\alpha$ and $M$.

4) Differential entropy: Let $X$, $Z_1$, and $Z_2$ be independent Gaussian random variables with mean zero and variances $\mathbb{E}[X^2] = P$ and $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_2^2] = N$. Let $Y_1 = g_1 X + Z_1$ and $Y_2 = g_2 X + Z_2$ for some constants $g_1, g_2 \in \mathbb{R}$. Express the following in terms of $P, N, g_1$, and $g_2$:
   a) $h(Z_1, Z_2)$.
   b) $h(Y_1, Y_2)$.
   c) $I(X; Y_1, Y_2)$. 
5) More differential entropy: Let $X, Y$ be jointly Gaussian with mean zero, variance one, and covariance $\rho \in (0, 1)$.
   a) What is $h(5X + 17)$
   b) What $h(X, Y)$
   c) What is $h(|X|)$?

6) Entropy and KL divergence: Let $\mathcal{X} = \{1, \ldots, L + M\}$, for $M, N \in \mathbb{N}$ and consider the distributions $P, Q \in \mathcal{P}(\mathcal{X})$ whose PMFs are, respectively,
   $p(x) = \begin{cases} p_x, & x = 1, \ldots, L; \\ 0, & x = L + 1, \ldots, L + M. \end{cases}$
   $q(x) = \begin{cases} \alpha p_x, & x = 1, \ldots, L; \\ \frac{1 - \alpha}{M}, & x = L + 1, \ldots, L + M. \end{cases}$

   where $0 < \alpha < 1$.
   (a) compute $D_{KL}(P \parallel Q)$
   (b) Express $H(Q)$ in terms of $H(P), \alpha$ and $M$.

7) Axiomatic definition of entropy: If a sequence of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ satisfies the following properties:
   - Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = 1$,
   - Continuity: $H_2(p, 1 - p)$ is a continuous function of $p$,
   - Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2)H_2(\frac{p_1}{p_1 + p_2}, \frac{p_3}{p_1 + p_2})$,

   prove that $H_m$ must be of the form
   $$H_m(p_1, p_2, \ldots, p_m) = - \sum_{i=1}^{m} p_i \log(p_i).$$

   Hint 1: Using induction show that
   $$H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + \ldots + p_k, p_{k+1}, \ldots, p_m) + (p_1 + \ldots + p_k)H_2(\frac{p_1}{p_1 + \ldots + p_k}, \ldots, \frac{p_k}{p_1 + \ldots + p_k})$$
   for all $k = 1, \ldots, m$.

   Hint 2: Let $f(m) = H_m(1/m, 1/m, \ldots, 1/m)$. Show that for two integers $i$ and $j$, it holds that $f(ij) = f(i) + f(j)$.

   Hint 3: Prove that $H_2(p, 1 - p) = -p \log p - (1 - p) \log(1 - p)$ for any rational $p$. Use continuity to extend the argument to real numbers.

8) Entropy under constraints: Let $X, Y, Z \sim \text{Ber}(1/2)$ and pairwise independent ($I(X; Y) = I(Y; Z) = I(X; Z) = 0$).
   a) Under this constraint, what is the minimum value for $H(X, Y, Z)$?
   b) Given an example achieving this minimum.

   Now instead of pairwise independence, assume that $I(X; Y) = I(Y; Z) = I(X; Z) = \alpha$ for some $\alpha \in [0, 1]$. Repeat parts (a) and (b).
9) **Directed information:** The *directed information* \( I(X^n \rightarrow Y^n) \) from \( X^n := (X_1, \ldots, X_n) \) to \( Y^n := (Y_1, \ldots, Y_n) \) (random correlated sequences) is an information measure that appears in the context of interactive communication and communication with feedback. It is defined as

\[
I(X^n \rightarrow Y^n) = \sum_{i=1}^n I(X_i; Y_i | Y^{i-1})
\]

That is, it is the sum of the the mutual information of inputs up to time \( i \) and the output at time \( i \) conditioned on the past outputs up to time \( i - 1 \). For this problem you can restrict yourself to considering discrete sources only (although this is not necessary).

(a) Show that \( I(X^n \rightarrow Y^n) \neq I(Y^n \rightarrow X^n) \) in general.

**Hint:** consider \( X^n \) and \( Y^n \) to be certain subsets of \( \{Z_0, Z_1, \ldots, Z_n\} \) that are i.i.d. Bern(1/2)).

(b) Consider a DMC used for communication with feedback. It is defined as

\[
\text{Codebook size } W > 0 \text{ and an input standard deviation parameter } \eta \in \mathbb{R}, \text{ and produces a sample of the (random) output sequence } (Y_1, \ldots, Y_n), \text{ for } Y_i \text{ as above.}
\]

Next implement a randomly generated Gaussian code. Let \( \text{Code}(n, W, \eta) \) be the function that takes as input a blocklength \( n \in \mathbb{N} \), a noise parameter \( \sigma \in \mathbb{R}_{>0} \) and an input sequence \( u \in \mathbb{R}^n \), and produces a sample of the (random) output sequence \( (Y_1, \ldots, Y_n) \) for \( Y_i \) as above.

b) Next implement a randomly generated Gaussian code. Let \( \text{Code}(n, W, \eta) \) be the function that takes as input the blocklength \( n \in \mathbb{N} \), a codebook size \( W \in \mathbb{N} \), and an input standard deviation parameter \( \eta \in \mathbb{R}_{>0} \), and outputs a collection \( \{u(w)\}_{w=1}^W \) of i.i.d. (across codewords and across time) sequences of length \( n \), where each symbol \( u_i(w) \), for \( i = 1, \ldots, n \) and \( w = 1, \ldots, W \), is drawn according to \( \mathcal{N}(0, \eta^2) \).

c) Show that the induced output probability density function \( q_Y : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0} \) is the Gaussian mixture model

\[
q_Y(v) = \frac{1}{W} \sum_{w=1}^W \varphi_\sigma(v - u(w)),
\]

where \( \varphi_\sigma(x) = \frac{1}{(2\pi \sigma^2)^{n/2}} e^{-\frac{|x|^2}{2\sigma^2}} \) is the density of \( \mathcal{N}(0, \sigma^2 I_n) \) and \( I_n \) is the \( n \times n \) identity matrix.

d) Let \( U \sim \mathcal{N}(0, \eta^2) \) be the coding variable and \( V = U + Z \) be the (single-letter) channel output, where \( Z \sim \mathcal{N}(0, \sigma^2) \) is independent of \( U \). Show that the target distribution for fixed \( \eta \) and \( \sigma \), i.e., the marginal distribution of \( V \) above, is \( \mathcal{N}(0, (\eta^2 + \sigma^2) I_n) \), and write out its probability density function \( \varphi_{\sqrt{\eta^2 + \sigma^2}}(v) \) explicitly.

e) Compute \( I(U, V) \) in terms of \( \eta \) and \( \sigma \).

f) Fix \( \eta = \sigma = 1, n \in \{1, 2\} \) (implement both cases) and let \( W \) range from 1 to \( 10^4 \) (choose appropriate gaps). For
each $W$ (and $n$), use the function $\text{Code}(n, W, \eta)$ to produce a Gaussian codebook. Compute and plot $q_V(v)$ from (3) versus $\varphi \sqrt{\eta^2 + \sigma^2}(v)$, for $v_i \in [-6, 6]$, $i = 1, \ldots, n$. Also plot the (scaled) conditional distributions $q_{V|W}(v|w) = \varphi \cdot (v - \mu_w)$, for $w = 1, \ldots, W$, on the same axes. Repeat this experiment for each considered $W$. Does the approximation of $\varphi \sqrt{\eta^2 + \sigma^2}$ via $q_V$ improves as $W$ grows? How do the results differ between the $n = 1$ and $n = 2$ cases?

g) Plot the total variation distance

$$
\delta_{TV}(q_V, \varphi \sqrt{\eta^2 + \sigma^2}) = \frac{1}{2} \int_{\mathbb{R}^n} |q_V(v) - \varphi \sqrt{\eta^2 + \sigma^2}(v)| dv
$$

versus the range of $W$ values. Describe and explain the curve you obtain.

11) **Multiple cascaded BSCs**: In this problem we study a generalization of the cascade of BSCs from Question 7 of Homework Sheet 5. Consider a cascade of $k$ identical and independent binary symmetric channels, each with crossover probability $\alpha$.

a) In the case where no encoding or decoding is allowed at the intermediate terminals, what is the capacity of this cascaded channel as a function of $k, \alpha$?

b) Now, assume that encoding and decoding is allowed at the intermediate points, what is the capacity as a function of $k, \alpha$?

c) What is the capacity of each of the above settings in the case where the number of cascaded channels, $k$, goes to infinity?

12) **Entropy power inequality**: A famous (and highly useful) information inequality is the **entropy power inequality (EPI)**. 

**Lemma (Entropy power inequality)** Let $X$ and $Y$ be two real-valued independent random variables. Then,

$$
e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}, \quad (4)
$$

with equality if and only if $X$ and $Y$ are jointly Gaussian.

Let us consider a special case of that result. Suppose $X$ and $Y$ are two independent random variables with density functions

$$
f_X(x) = \begin{cases} 
\frac{1}{2a} & |x| \leq a, \\
0 & |x| > a
\end{cases}
$$

and

$$
f_Y(y) = \begin{cases} 
\frac{1}{2b} & |y| \leq b, \\
0 & |y| > b
\end{cases}
$$

for some arbitrary $0 < a \leq b$.

a) Compute $h(X)$ and $h(Y)$.

b) Find the probability density function of $Z = X + Y$. You may solve analytically or rely on a carefully labeled graphical solution.

c) Find $h(Z)$. 

Hint: For $\beta \geq \alpha$, we have

$$\int_{\alpha}^{\beta} x \log x dx = \frac{1}{2} \beta^2 \log \beta - \frac{1}{2} \alpha^2 \log \alpha - \frac{\log e}{4} (\beta^2 - \alpha^2).$$

13) Erasures and errors in a binary channel: Consider a binary channel with probability of error $\alpha$ and probability of erasure $\epsilon$ as depicted in Figure 1. More specifically, consider $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ where $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, \epsilon\}$ and $P_{Y|X}$ described by the relation:

$$Y = \begin{cases} X, & \text{w.p. } 1 - \alpha - \epsilon, \\ 1 - X, & \text{w.p. } \alpha, \\ \epsilon, & \text{w.p. } \epsilon. \end{cases}$$

Find a closed from expression for the capacity $\max_{P_X} I(X; Y)$ of this channel.

![Fig. 1: Erasures and errors in a binary channel](image)

14) Modulus channel: Consider a discrete channel with input alphabet $\mathcal{X} = \{0, 1, \ldots, q - 1\}$. The channel output is

$$Y = [X + Z] \mod q$$

where $Z$ is independent of $X$ with $p_Z(0) = 1 - \beta$ and $p_Z(z) = \frac{\beta}{q-1}$ for $z = 1, 2, \ldots, q - 1$.

a) What is $H(Z)$?

b) What is the capacity of this channel?

15) Time varying channels: Consider a time varying binary symmetric channel. More specifically, at time $i = 1, \ldots, n$, the channel is specified by $(\mathcal{X}, \mathcal{Y}, P_{Y_i|X_i})$, where $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $P_{Y_i|X_i}$ is described by the relationship $Y_i = X_i \oplus Z_i$, where $Z_i \sim \text{Bern}(p_i)$ with $p_i \in (0, 1)$. Assume that $\{Z_i\}_{i=1}^n$ are independent, and, thus, $Y_i$’s are conditionally independent given $X_i$’s. Find $\max_{P_X^n} I(X^n; Y^n)$, where the underlying distribution is $P_X \prod_{i=1}^n P_{Y_i|X_i}$. 
16) **Computing channel capacity:** Consider a channel \((\mathcal{X}, \mathcal{Y}, P_{Y|X})\), where \(\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}\) and \(P_{Y|X}\) has a conditional PMF \(p_{Y|X}\) given by

\[
p_{Y|X} = \begin{bmatrix}
\frac{2}{3} & \frac{1}{3} & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
0 & \frac{1}{3} & \frac{2}{3}
\end{bmatrix}
\]

a) Find the capacity \(\max_{p_X} I(X; Y)\) and the distribution that achieves it.

b) Qualitatively justify why the distribution found in part (a) achieves the capacity.