1. Proofs
Consider $f(x), g(x)$ two probability density functions and $X$ a continuous random variable. Prove that:

(a) $h(X + a) = h(X)$ for any constant $a$.
(b) $h(aX) = h(X) + \log |a|$, for any constant $a \neq 0$.
(c) $D(f \| g) \geq 0$, with equality if and only if $f = g$.
(d) $I(X; Y)$ for $X, Y$ with a joint density, which we defined as the relative entropy between the joint density and the product densities, is $I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y)$.

Solutions: Proofs

(a) Let $Y = X + a$. Then $f_Y(y) = f_X(y - a)$ since the densities are translation invariant. Therefore, $h(X + a) = h(X)$.

(b) Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(y/a)$, and

$$h(aX) = - \int f_Y(y) \log f_Y(y) dy$$

$$= - \int \frac{1}{|a|} f_X(\frac{y}{a}) \log \left( \frac{1}{|a|} f_X(\frac{y}{a}) \right) dy$$

$$= - \int f_X(x) \log(f_X(x)) dx + \log |a|$$

$$= h(X) + \log |a|$$
(c) Let \( S \) be the support set of \( f \). Then

\[
-D(f \| g) = - \int_S f(x) \log \frac{f(x)}{g(x)} \, dx
\]

(5)

\[
= \int_S f(x) \log \frac{g(x)}{f(x)} \, dx
\]

(6)

\[
\leq \log \int_S f(x) \frac{g(x)}{f(x)} \, dx
\]

(7)

\[
= \log \int_S g \, dx
\]

(8)

\[
\leq \log 1
\]

(9)

\[
= 0
\]

(10)

where (??) follows by Jensen’s inequality. Therefore \( D(f \| g) \geq 0 \).

(d) \[
I(X;Y) = D(f_{X,Y} \| f_X f_Y)
\]

\[
= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dxdy
\]

\[
= \int f_{X,Y}(x,y) \log \frac{f_{X|Y}(x|y)}{f_X(x)} \, dxdy
\]

(11)

\[
= \int f_{X,Y}(x,y) \log f_{X|Y}(x|y) \, dxdy - \int f_{X,Y}(x,y) \log f_X(x) \, dxdy
\]

(12)

\[
= \int f_{X,Y}(x,y) \log f_{X|Y}(x|y) \, dxdy - \int f_X(x) \log f_X(x) \, dx
\]

(13)

\[
= h(X) - h(X|Y).
\]

Similarly, we can get \( I(X;Y) = h(Y) - h(Y|X) \). Also,

\[
I(X;Y) = D(f_{X,Y} \| f_X f_Y)
\]

\[
= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dxdy
\]

\[
= \int f_{X,Y}(x,y) \log f_{X|Y}(x,y) \, dxdy
\]

\[
- \int f_{X,Y}(x,y) \log f_X(x) \, dxdy - \int f_{X,Y}(x,y) \log f_Y(y) \, dxdy
\]

\[
= \int f_{X,Y}(x,y) \log f_{X|Y}(x,Y) \, dxdy - \int f_X(x) \log f_X(x) \, dx - \int f_Y(y) \log f_Y(y) \, dy
\]

\[
= h(X) + h(Y) - h(X,Y)
\]
2. Differential entropy.
Evaluate the differential entropy \( h(X) = -\int f \ln f \) for the following:

(a) The exponential density, \( f(x) = \lambda e^{-\lambda x}, \; x \geq 0 \).
(b) The Laplace density, \( f(x) = \frac{1}{2\lambda} e^{-\lambda|x|} \).
(c) The sum of \( X_1 \) and \( X_2 \), where \( X_1 \) and \( X_2 \) are independent normal random variables with means \( \mu_i \) and variances \( \sigma_i^2, \; i = 1, 2 \).

**Solution: Differential Entropy.**

(a) Exponential distribution.
\[
h(f) = - \int_0^{\infty} \lambda e^{-\lambda x}[\ln \lambda - \lambda x]dx
= -\ln \lambda + 1 \text{ nats.}
= \log \frac{e}{\lambda} \text{ bits.}
\]

(b) Laplace density.
\[
h(f) = - \int_{-\infty}^{\infty} \frac{1}{2\lambda} e^{-\lambda|x|}[\ln \frac{1}{2} + \ln \lambda - \lambda|x|]dx
= -\ln \frac{1}{2} - \ln \lambda + 1
= \ln \frac{2e}{\lambda} \text{ nats.}
= \log \frac{2e}{\lambda} \text{ bits.}
\]

(c) Sum of two normal distributions.
The sum of two normal random variables is also normal, so applying the result derived the class for the normal distribution, since \( X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \),
\[
h(f) = \frac{1}{2} \log 2\pi e(\sigma_1^2 + \sigma_2^2) \text{ bits.}
\]

3. Mutual information for correlated normals.
Find the mutual information \( I(X; Y) \), where
\[
\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left( 0, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \right), \quad -1 < \rho < 1.
\]
Evaluate \( I(X; Y) \) for \( \rho = 0 \), and the limit of \( I(X; Y) \) as \( \rho \) approaches 1. Comment on your findings.
Solution: Mutual information for correlated normals.

\[
\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}_2 \left( 0, \begin{bmatrix} \sigma^2 & \rho \sigma^2 \\ \rho \sigma^2 & \sigma^2 \end{bmatrix} \right)
\]

Using the expression for the entropy of a multivariate normal

\[ h(X, Y) = \frac{1}{2} \log(2\pi e)^2 |K| = \frac{1}{2} \log(2\pi e)^2 \sigma^4 (1 - \rho^2). \]

Since \( X \) and \( Y \) are individually normal with variance \( \sigma^2 \),

\[ h(X) = h(Y) = \frac{1}{2} \log 2 \pi e \sigma^2. \]

Hence

\[ I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log(1 - \rho^2). \]

(a) \( \rho = 0 \). In this case, \( X \) and \( Y \) are independent, and hence \( I(X; Y) = 0 \), which agrees with the formula.

(b) It is easy to see that \( \lim_{\rho \to 1} -\frac{1}{2} \log(1 - \rho^2) = \infty \). This is reasonable since \( \rho = 1 \) implies \( X = Y \), and knowing \( X \) implies perfect knowledge about \( Y \). Hence the mutual information should be infinite, which agrees with the formula.

4. Postprocessing the output.
One is given a communication channel with transition probabilities \( p(y | x) \) and channel capacity \( C = \max_{p(x)} I(X; Y) \). A helpful statistician postprocesses the output by forming \( \tilde{Y} = g(Y) \), yielding a channel \( p(\tilde{y}|x) \). He claims that this will strictly improve the capacity. Show that he is wrong.

Solution: Preprocessing the output.
The statistician calculates \( \tilde{Y} = g(Y) \). Since \( X \to Y \to \tilde{Y} \) forms a Markov chain, we can apply the data processing inequality. Hence for every distribution on \( x \),

\[ I(X; Y) \geq I(X; \tilde{Y}). \]

Let \( \tilde{p}(x) \) be the distribution on \( x \) that maximizes \( I(X; \tilde{Y}) \). Then

\[ C = \max_{p(x)} I(X; Y) \geq I(X; Y)_{p(x) = \tilde{p}(x)} \geq I(X; \tilde{Y})_{p(x) = \tilde{p}(x)} = \max_{p(x)} I(X; \tilde{Y}) = \tilde{C}. \]

Thus, the helpful suggestion is wrong and processing the output does not increase capacity.

5. The Z channel.
The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \quad x, y \in \{0, 1\}$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

**Solution: The Z channel.**

First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $\alpha = \Pr(X = 1)$:

$$H(Y|X) = \Pr(X = 0) \cdot 0 + \Pr(X = 1) \cdot 1 = \alpha$$

$$H(Y) = H(\Pr(Y = 1)) = H(\alpha/2)$$

$$I(X;Y) = H(Y) - H(Y|X) = H(\alpha/2) - \alpha$$

Since $I(X;Y)$ is strictly concave on $\alpha$ (why?) and $I(X;Y) = 0$ when $\alpha = 0$ and $\alpha = 1$, the maximum mutual information is obtained for some value of $\alpha$ such that $0 < \alpha < 1$.

Using elementary calculus, we determine that

$$\frac{d}{d\alpha} I(X;Y) = \frac{1}{2} \log_2 \frac{1 - \alpha/2}{\alpha/2} - 1,$$

which is equal to zero for $\alpha = 2/5$. (It is reasonable that $\Pr(X = 1) < 1/2$ since $X = 1$ is the noisy input to the channel.) So the capacity of the Z-channel in bits is $H(1/5) - 2/5 = 0.722 - 0.4 = 0.322$.

6. Choice of channels.

Find the capacity $C$ of the union of 2 channels $(X_1, p_1(y_1|x_1), Y_1)$ and $(X_2, p_2(y_2|x_2), Y_2)$ where, at each time, one can send a symbol over channel 1 or over channel 2 but not
both. Assume the output alphabets are distinct and do not intersect. Show that 
\[ 2^C = 2^{C_1} + 2^{C_2}. \]

**Solution: Choice of channels.**

This means that \( 2^C \) is the effective alphabet size of a channel with capacity \( C \). Consider

the following communication scheme:

\[
X = \begin{cases} 
X_1 & \text{Probability } \alpha \\
X_2 & \text{Probability } (1 - \alpha)
\end{cases}
\]

Let

\[
\theta(X) = \begin{cases} 
1 & X = X_1 \\
2 & X = X_2
\end{cases}
\]

Since the output alphabets \( \mathcal{Y}_1 \) and \( \mathcal{Y}_2 \) are disjoint, \( \theta \) is a function of \( Y \) as well, i.e. \( X \rightarrow Y \rightarrow \theta \).

\[
I(X;Y,\theta) = I(X;\theta) + I(X;Y|\theta)
\]

\[
= I(X;Y) + I(X;\theta|Y)
\]

Since \( X \rightarrow Y \rightarrow \theta \), \( I(X;\theta|Y) = 0 \). Therefore,

\[
I(X;Y) = I(X;\theta) + I(X;Y|\theta)
\]

\[
= H(\theta) - H(\theta|X) + \alpha I(X_1;Y_1) + (1 - \alpha)I(X_2;Y_2)
\]

\[
= H(\alpha) + \alpha I(X_1;Y_1) + (1 - \alpha)I(X_2;Y_2)
\]

Thus, it follows that

\[
C = \sup_\alpha \{ H(\alpha) + \alpha C_1 + (1 - \alpha)C_2 \}.
\]

Maximizing over \( \alpha \) one gets the desired result. The maximum occurs for \( H'(\alpha) + C_1 - C_2 = 0 \), or \( \alpha = 2^{C_1}/(2^{C_1} + 2^{C_2}) \).

7. **Channel capacity.**

Find the capacity of the following channels with probability transition matrices:

(a) \( \mathcal{X} = \mathcal{Y} = \{0, 1, 2\} \)

\[
p(y|x) = \begin{bmatrix} 
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3
\end{bmatrix}
\]
(b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

(c) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}$

$$p(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix}$$

(d) Calculate the capacity of the following channel:

```
<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1-p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1-p</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Solution: Channel capacity.

(a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

This is a symmetric channel and by the known result for symmetric channel:

$$C = \log |\mathcal{Y}| - H(\mathbf{r}) = \log 3 - \log 3 = 0.$$ 

In this case, the output is independent of the input.

(b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$
Again the channel is symmetric:

\[ C = \log |\mathcal{Y}| - H(\mathbf{r}) = \log 3 - \log 2 = 0.58 \text{ bits} \]

(c) \( \mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\} \)

\[
p(y|x) = \begin{bmatrix}
p & 1 - p & 0 & 0 \\
1 - p & p & 0 & 0 \\
0 & 0 & q & 1 - q \\
0 & 0 & 1 - q & q \\
\end{bmatrix}
\]

This channel consists of a sum of two BSC’s, and using the result of problem 1, the capacity of the channel is

\[ C = \log \left( 2^{1 - H(p)} + 2^{1 - H(q)} \right) \]

(d) This channel consists of a sum of a BSC and a zero-capacity channel. Thus

\[ C = \log \left( 2^{1 - H(p)} + 1 \right) \]