1. Each of the following lists two quantities. Determine (i) whether $\leq$ or $\geq$ holds between them, and (ii) under what conditions they are equal. Don't forget to justify your answers. Assume X, Y are random variables, $f(X)$ is some function of X, and a_i for $i = 1, \ldots, n$ are real numbers.

   (a) $H(X,Y)$ and $H(X) + H(Y)$

   (b) $H(X,Y)$ and $H(X)$

   (c) $\max_{i \in \{1,\ldots,n\}} a_i$ and $\frac{1}{n}\sum_{i=1}^{n} a_i$

   (d) $I(X;Y|f(X))$ and $I(X;Y)$

a) $H(X,Y) = H(X) + H(Y) - I(X;Y)$

$\leq H(X) + H(Y)$

Equality holds iff $I(X;Y) = 0$, i.e. X and Y are independent

b) $H(X,Y) = H(X) + H(Y|X)$

$\geq H(X)$

Equality holds iff $H(Y|X) = 0$, i.e. $Y = g(X)$ for some function g

c) Let $a^* = \max_i a_i$

$$\frac{1}{n} \sum_{i=1}^{n} a_i \leq \frac{1}{n} \sum_{i=1}^{n} a^* = a^*$$

Thus $\geq$ holds.

Equality holds iff $a_i = a^*$ for all i, i.e. all a_i are equal

d) $$I(X;Y) = I(X, f(X) ; Y)$$

$$= I(f(X); Y) + I(X; Y | f(X))$$

$$\geq I(X;Y| f(X))$$

Thus $\leq$ holds.

Equality holds iff $I(f(X); Y) = 0$, i.e. $f(X)$ and Y are independent

2. This problem illustrates that any entropy can be written in terms of the binary entropy function. Recall that the binary entropy function is given by

$$H(p) = -p \log p - (1-p)\log(1-p).$$

Let $X \in \{1,2,3\}$ be a random variable with probabilities (p_1, p_2, p_3). Let

$$Y = \begin{cases} 1 & \text{if } X = 1 \\ 0 & \text{if } X = 2 \text{ or } X = 3. \end{cases}$$

(a) Prove that $H(X) = H(Y) + (p_2+p_3)H(X|Y=0)$. *Hint:* First show that $H(X) = H(X,Y)$.

(b) Given part (a), write $H(X)$ in terms of p_1, p_2, p_3 using only the binary entropy function.

(c) Now consider the random variable $Z \in \{1,2,3,4\}$ with probabilities (p_1, p_2, p_3, p_4). Using a similar method as above, write $H(Z)$ using only the binary entropy function.

a) Since Y is a function of X, $H(Y|X) = 0$. so

$$H(X) = H(X) + H(Y|X)$$

=0 because $X=1$ if $Y=1$

$$= H(X,Y)$$

$$= H(Y) + H(X|Y)$$

$$= H(Y) + p_1 H(X|Y=1) + (p_2+p_3)H(X|Y=0)$$

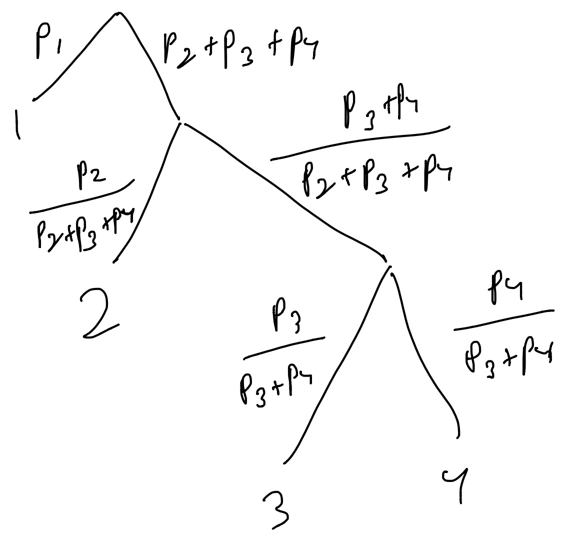
$$= H(Y) + (p_2+p_3)H(X|Y=0)$$

b) Note $H(Y) = H(p_1)$

$$H(X|Y=0) = H\left(\frac{p_2}{p_2+p_3}\right)$$

Thus $$H(X) = H(p_1) + (p_2+p_3)H\left(\frac{p_2}{p_2+p_3}\right)$$

c) We can construct Z using a binary tree as follows:



Thus

$$H(Z) = H(p_1) + (p_2+p_3+p_4)\, H\left(\frac{p_2}{p_2+p_3+p_4}\right)$$
$$+ (p_3+p_4)\, H\left(\frac{p_3}{p_3+p_4}\right)$$

3. Consider a random variable X with alphabet $\mathcal{X}$ and distribution $p(x)$. Given real numbers a, b, define the set

$$\mathcal{S} = \{x \in \mathcal{X} : a \leq p(x) \leq b\}.$$

Prove the following:

$$\frac{\Pr\{X \in \mathcal{S}\}}{b} \leq |\mathcal{S}| \leq \frac{\Pr\{X \in \mathcal{S}\}}{a}.$$

Since $p(x) \geq a$ for all $x \in S$:

$$\Pr\{X \in S\} = \sum_{x \in S} p(x)$$

$$\geq \sum_{x \in S} a$$

$$= |S| a \Rightarrow |S| \leq \frac{\Pr\{X \in S\}}{a}$$

Since $p(x) \leq b$ for all $x \in S$:

$$\Pr\{X \in S\} = \sum_{x \in S} p(x)$$

$$\leq \sum_{x \in S} b$$

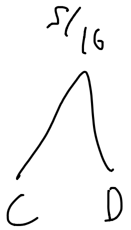
$$= |S| b \Rightarrow |S| \geq \frac{\Pr\{X \in S\}}{b}$$

4. Consider a distribution on four letters $\{A, B, C, D\}$ with probabilities 3/8, 5/16, 1/4, 1/16.

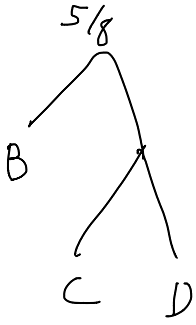
(a) Find a Huffman code for this distribution. (Not just the code-lengths. Give a specific code.)

(b) Find a Shannon code for this distribution. Recall that a Shannon code is a *prefix* code with code lengths given by $\lceil \log \frac{1}{p(x)} \rceil$.

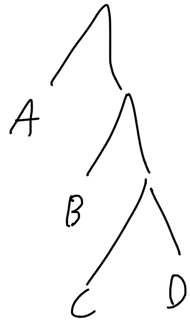
a) Merge C & D:



Merge B & (C,D):



Merge A&(B,C,D):



Example code:

A	0
B	10
C	110
D	111

b) Lengths should be:

A = 2

B = 2

C = 2

D = 4

Example code:

A	00
B	01
C	10
D	1100

5. The m-ary symmetric channel is a generalization of the binary symmetric channel. The input and output alphabets each consist of the integers from 1 to m; i.e.

$$\mathcal{X} = \mathcal{Y} = \{1, 2, \ldots, m\}.$$

The channel transition matrix is given by

$$p(y|x) = \begin{cases} 1-p & \text{if } y = x \\ \frac{p}{m-1} & \text{if } y \neq x. \end{cases}$$

(a) For any distribution on X, calculate $H(Y|X)$ for the m-ary symmetric channel.

(b) Find the capacity of the m-ary symmetric channel.

(c) Use Fano's inequality to find an upper bound on $H(Y|X)$ using the fact that $p = \Pr\{Y \neq X\}$. Compare this bound to the answer to part (a).

a) $$H(Y|X) = H\Big(1-p, \underbrace{\frac{p}{m-1}, \frac{p}{m-1}, \ldots, \frac{p}{m-1}}_{m-1 \text{ times}}\Big)$$

$$= -(1-p)\log(1-p) - (m-1)\cdot\frac{p}{m-1}\log\left(\frac{p}{m-1}\right)$$

$$= -(1-p)\log(1-p) - p\log p + p\log(m-1)$$

$$= H(p) + p\log(m-1)$$

b) $$I(X;Y) = H(Y) - H(Y|X)$$

$$= H(Y) - H(p) - p\log(m-1)$$

$$\leq \log m - H(p) - p\log(m-1)$$

Upper bound achievable if $X \sim \text{Unif}\{1, \ldots, m\}$

Thus $C = \log m - H(p) - p\log(m-1)$

c) By Fano's inequality, where $P_e = \Pr(Y \neq X) = p$

$$H(Y|X) \leq H(P_e) + P_e \log(|\mathcal{Y}|-1)$$
$$= H(p) + p \log(m-1)$$

This is exactly the same as the quantity from (a).

Thus the m-ary symmetric channel is precisely the conditional distribution that achieves equality in Fano's inequality.

6. Consider the following source code on alphabet $\{a, b, c, d, e\}$ which is *not* a prefix code

$$\begin{aligned} a &\to 0\,0 \\ b &\to 0\,1 \\ c &\to 1\,1 \\ d &\to 0\,0\,1 \\ e &\to 0\,1\,1 \end{aligned}$$

(a) Even though this is not a prefix code, it is still decodable. Decode the following bit string, which represents the concatenation of several codewords:

$$0\,0\,1\,1\,1\,0\,1\,0\,1\,1\,0\,0\,1\,1$$

(b) Find a prefix code that is equivalent to this code in that it has the same codeword lengths.

(c) Describe the advantage of a prefix code.

Solution:

(a) The bit string is decoded as $d\,c\,b\,e\,a\,c$.

(b) There are many prefix codes with the same codeword lengths, for example:

$$\begin{aligned} a &\to 0\,0 \\ b &\to 0\,1 \\ c &\to 1\,0 \\ d &\to 1\,1\,0 \\ e &\to 1\,1\,1 \end{aligned}$$

(c) In addition to always being uniquely decodable, prefix codes are instantaneously decodable. This means that the boundaries between codewords can be immediately be identified by parsing the string from left to right. That is, any point that *could* be the end of a codeword always *is* the end of a codeword, since no codeword is a prefix of another codeword. For example, in part (a) the initial part of the string $0\,0\,1$ could have either been the codeword for a followed by the beginning of the codeword for c, or the codeword for d; it was impossible to tell without reading further ahead in the string. This does not occur with a prefix code.

7. Consider the steps of the converse proof to the channel coding theorem shown below. Justify each of the steps (a)–(e).

$$
\begin{aligned}
nR &\overset{\text{(a)}}{\leq} I(W;Y^n) + n\epsilon_n \\
&\overset{\text{(b)}}{\leq} I(X^n;Y^n) + n\epsilon_n \\
&\overset{\text{(c)}}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&\overset{\text{(d)}}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&\overset{\text{(e)}}{\leq} nC^{(I)} + n\epsilon_n
\end{aligned}
$$

where $\epsilon_n = \frac{1}{n} + P_e^{(n)}R$.

Solution:

(a) Since the message W is uniformly distributed from the set $\{1, \ldots, 2^{nR}\}$, $nR = H(W)$. By the definition of mutual information, $H(W) = I(W;Y^n) + H(W|Y^n)$. By Fano's inequality, $H(W|Y^n) \leq 1 + P_e^{(n)} nR = n\epsilon_n$. Putting this together gives $nR \leq I(W;Y^n) + n\epsilon_n$.

(b) Since $W \rightarrow X^n \rightarrow Y^n$ is a Markov chain, by the data processing inequality $I(W;Y^n) \leq I(X^n;Y^n)$.

(c) By the definition of mutual information, $I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n)$. Applying the chain rule we have $H(Y^n|X^n) = \sum_{i=1}^n H(Y_i|X^n, Y_1, \ldots, Y_{i-1})$. By the memorylessness of the channel, Y_i depends only on X_i, so $H(Y^n|X^n) = \sum_{i=1}^n H(Y_i|X_i)$. This proves $I(X^n;Y^n) = H(Y^n) - \sum_{i=1}^n H(Y_i|X_i)$.

(d) By the independence bound $H(Y^n) \leq \sum_{i=1}^n H(Y_i)$.

(e) By the definition of mutual information we have

$$\sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i;Y_i).$$

Since by definition $C^{(I)} = \max_{p(x)} I(X;Y)$, it must be that $I(X_i;Y_i) \leq C^{(I)}$ for each i. Thus the above sum is upper bounded by $nC^{(I)}$.

8. Consider the following channel. The input $X \in \{0, 1\}$ is sent through two independent binary erasure channels, each with erasure probability α. Let Y_a and Y_b be the two outputs. The output of the channel is the pair (Y_a, Y_b). Find the capacity of this channel.

Solution: If the input to the channel is 0, then the output distribution is

$$\begin{cases} 00, & \text{w.p. } (1-\alpha)^2 \\ 0e, & \text{w.p. } \alpha(1-\alpha) \\ e0, & \text{w.p. } \alpha(1-\alpha) \\ ee, & \text{w.p. } \alpha^2 \end{cases}$$

and a similar distribution occurs if the input is 1. Thus, the input can be determined exactly from the output in every case except when $Y = ee$, i.e. when both erasure channels have an erasure, which occurs with probability α^2. Thus, this channel acts like an erasure channel with erasure probability α^2, where outputs 00, $0e$, $e0$, 11, $1e$, $e1$ all count as "no erasure", in that the input can be determined from the output. By this argument, the capacity is $1 - \alpha^2$.

For a somewhat more rigorous argument, we may do the following. Let E be a random variable equal to 1 if $Y = ee$ and 0 otherwise. Then we may write

$$\begin{aligned} I(X;Y) &\overset{\text{(a)}}{=} I(X;Y,E) \\ &\overset{\text{(b)}}{=} I(X;E) + I(X;Y|E) \\ &\overset{\text{(c)}}{=} \Pr\{E=0\}I(X;Y|E=0) + \Pr\{E=1\}I(X;Y|E=1) \\ &\overset{\text{(d)}}{=} (1-\alpha^2)H(X) \\ &\overset{\text{(e)}}{\leq} 1-\alpha^2 \end{aligned}$$

where

- (a) follows because E is a function of Y,
- (b) follows from the chain rule,
- (c) follows because X and E are independent, and from expanding the mutual information,
- (d) follows because $\Pr\{E = 0\} = 1 - \alpha^2$, because when $E = 1$ then $Y = ee$ and thus fixed, and because if $E = 0$ then X and Y determine each other,
- (e) follows by the uniform bound.

Equality occurs in the above if X is uniform on $\{0, 1\}$, so the capacity is $1 - \alpha^2$.

9. Let $p(x)$ be the uniform distribution on the finite alphabet $\mathcal{X}$, and $A_\epsilon^{(n)}$ the typical set with respect to $p(x)$. Prove that every sequence $x^n \in \mathcal{X}^n$ is typical.

Solution: For a uniform distribution $H(X) = \log |\mathcal{X}|$. Also for all $x^n \in \mathcal{X}^n$, $p(x^n) = |\mathcal{X}|^{-n}$. Recal that the typical set is given by

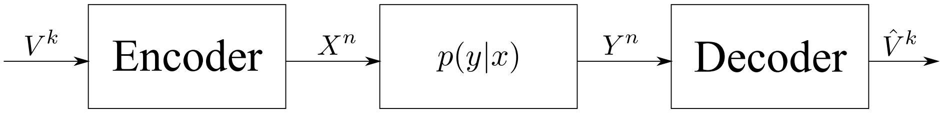
$$A_\epsilon^{(n)} = \left\{ x^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon \right\}.$$

For any x^n, we have

$$\begin{aligned} \left| -\frac{1}{n} \log p(x^n) - H(X) \right| &= \left| -\frac{1}{n} \log |\mathcal{X}|^{-n} - \log |\mathcal{X}| \right| \\ &= \big| \log |\mathcal{X}| - \log |\mathcal{X}| \big| = 0. \end{aligned}$$

Thus any sequence satisfies the condition for typicality for any ϵ, so all sequences are typical.

10. Consider the following joint source-channel coding problem in which the sequence V^k is to be transmitted across a channel. Note that the length of the source sequence is k and the number of channel uses is n.



Assume V^k is drawn from an i.i.d. distribution Bern(q) (i.e., 0 with probability $1 - q$ and 1 with probability q), and assume the channel is a binary symmetric channel with crossover probability p.

(a) Under what circumstances can $\Pr\{\hat{V}^k \neq V^k\}$ be made arbitrarily small?

(b) Suppose $n/k = m$ is an integer, and the encoder uses a repetition strategy. That is, it chooses

$$X^n = (\underbrace{V_1, V_1, \ldots, V_1}_{m \text{ times}}, \underbrace{V_2, V_2, \ldots, V_2}_{m \text{ times}}, \ldots, \underbrace{V_k, V_k, \ldots, V_k}_{m \text{ times}}).$$

The decoder chooses $\hat{V}_i$ as whichever of 0 or 1 appeared more often among the m bits associated with V_i. For example, if $m = 3$ and $k = 2$, then the channel output string $Y^n = (0, 0, 1, 1, 0, 1)$ would be decoded to $\hat{V}^k = (0, 1)$. In the case that $m = 3$, find $\Pr\{\hat{V}^k \neq V^k\}$.

(c) Now consider an arbitrary but fixed m. Does the repetition strategy result in arbitrarily small probability of error by taking $k \to \infty$? What about for fixed k and $m \to \infty$?

Solution:

(a) By the source-channel separation theorem, V^k can be reproduced with small probability of error if $k/n < C/H(V)$. In this case, $H(V) = H(q)$ and $C = 1 - H(p)$. Thus $\Pr\{V^k \neq \hat{V}^k\}$ can be made arbitrarily small if $k/n < \frac{1-H(p)}{H(q)}$.

(b) For a given V_i, the estimate $\hat{V}_i$ will be correct if a majority of the copies of V_i are not flipped by the channel. Since $m = 3$ copies are sent, this occurs if 0 or 1 copies are flipped, which happens with probability $(1-p)^3 + 3p(1-p)^2$. Thus $\Pr\{\hat{V}_i = V_i\} = (1-p)^3 + 3p(1-p)^2$. Since the source sequence and the channel operation are i.i.d., $\Pr\{\hat{V}^k = V^k\} = [(1-p)^3 + 3p(1-p)^2]^k$. Therefore $\Pr\{\hat{V}^k \neq V^k\} = 1 - [(1-p)^3 + 3p(1-p)^2]^k$.

(c) For fixed m and $k \to \infty$, each individual V_i is decoded with some positive probability of error (in the above case it was $1 - (1-p)^3 - 3p(1-p)^2$). No matter how small this probability is for each V_i, for large k, with high probability, at least one error will occur, meaning the probability of error will always go to 1.

For fixed k and $m \to \infty$, then the probability of error for each individual V_i vanishes as m grows. Since k remains fixed, the total probability of error becomes arbitrarily small as $m \to \infty$.