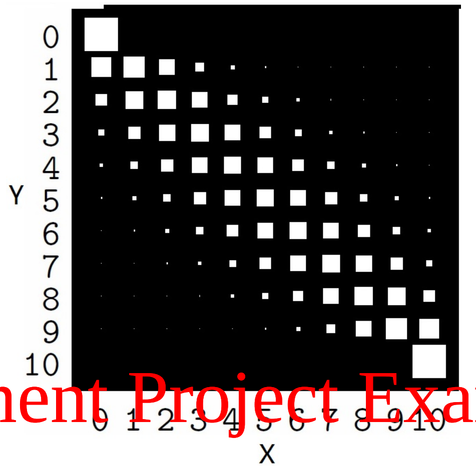
SECTION A.

Answer each of the following questions [Marks per questions as shown; 25% total]

1. (4 points) Suppose X and Y are random variables with outcomes $\{0, 1, \ldots, 10\}$. The figure below shows a *Hinton diagram* of the joint distribution of X and Y, with the area of the white squares at the cell for (x, y) proportional to the corresponding $P(X = x, Y = y)$. (Wherever there is no visible white square at the cell for (x, y), $P(X = x, Y = y)$ is close to 0.)

   Is it possible for X and Y to be statistically independent? Justify your answer.



No. If they were in $x) = p(Y = y)$. That is, the distribut ue. But this is evidently not so from the diagram, e.g. looking at the first and second c

2. (6 points) Let X and Y be two random variables s $\{0, 1\}$.

   (a) (2 pts) Does specifying $P(X|Y)$ and $P(Y$ $)$?

   Yes, because $P(X, Y) = P(X|Y)P(Y)$ by definition of joint probability.

   (b) (4 pts) Does specifying $P(X|Y)$ and $P(Y|X)$ fully determine $P(X, Y)$?

   Yes, because if we know $P(X|Y)$ and $P(Y|X)$, we can work out $P(Y)$ by inverting Bayes' rule. In particular we have $\frac{p(X|Y=1)}{p(X|Y=0)} \cdot \frac{p(Y=1)}{p(Y=0)} = \frac{p(Y=1|X)}{p(Y=0|X)}$. Note that this crucially exploits the simple structure of binary random variables.

   For both questions, if your answer is yes, then write down the formula of computing $P(X, Y)$ from the given quantities. Otherwise, give a counter-example.

3. (15 points) The Journal of Information Theory Research reviews all submitted papers in the following process. An editor first assigns the paper to two reviewers, who make the recommendation of either acceptance (1) or rejection (0). Based on their reviews, the editor makes the final decision. Even if the two reviews are consistent, the editor may still make the opposite decision. Let Z be the editor's decision, and X and Y be the reviewers' recommendation. Assume X and Y are independent and $P(X = 1) = P(Y = 1) = 0.5$. The conditional probability $P(Z = 0|X, Y)$ is given by

   (a) (6 pts) Compute $P(X = 1|Z = 0)$, showing all your working.

   We have $P(X = 1|Z = 0) = P(Z = 0|X = 1)P(X = 1)/P(Z = 0)$. Now $P(Z = 0|X = 1) = P(Z = 0|X = 1, Y = 0)P(Y = 0) + P(Z = 0|X =$

$P(Z=0 \mid X,Y)$	$X=0$	$X=1$
$Y=0$	0.9	0.5
$Y=1$	0.5	0.1

$1, Y = 1)P(Y = 1)$, which is $0.25 + 0.05 = 0.30$. Similarly $P(Z = 0|X = 0) = 0.45 + 0.25 = 0.70$. So, $P(Z = 0) = 0.5$. Thus $P(X = 1|Z = 0) = 0.30$.

(b) (2 pts) If from the above you find $P(X = 1|Z = 0) > P(X = 1)$, explain intuitively why the probability increased; else, if you find $P(X = 1|Z = 0) < P(X = 1)$, explain why the probability decreased.

It decreased. The reason is that it is unlikely that X voted yes, since that has a smaller chance of leading to $Z = 0$.

(c) (5 pts) Compute $P(X = 1|Z = 0, Y = 1)$, showing all your working.

This is $P(Z = 0|X = 1, Y = 1)P(X = 1|Y = 1)/P(Z = 0|Y = 1)$. The denominator is $P(Z = 0|X = 1, Y = 1)P(X = 1|Y = 1) + P(Z = 0|X = 0, Y = 1)P(X = 0|Y = 1) = 0.30$. Thus it is $(0.1)(0.5)/0.3 = 5/30 = 1/6 = 0.166...$.

(d) (2 pts) If you find $P(X = 1|Z = 0, Y = 1) > P(X = 1|Z = 0)$, explain intuitively why the probability increased; else, explain why the probability decreased.

It decreased. If we know that the editor voted to reject, we have some belief how likely it is that X voted for acceptance (30% chance from (a)). If we further learn that Y voted for acceptance, it seems unlikely that X also voted for acceptance, since the editor only very rarely overturns two recommendations for acceptance i.e. $p(Z = 0$

SECTION B.

Answer each of the following questions [Marks per questions as shown; 25% total]

1. (5 points) Let X, Y be two random variables with the following joint distribution:

$P(X,Y)$		$X=1$	$X=2$
Y	1	1/4	0
	2	1/4	1/2

   (a) (3 pts) Compute $H(Y|X)$.

   We have $p(X = 1) = 1/2$, and $p(X = 2) = 1/2$. Also, $p(Y = 1|X = 1) = 1/2$ and $p(Y = 1|X = 2) = 0$. So, $H(Y|X) = 0.5 \cdot (H(1/2) + H(0)) = 0.5$.

   (b) (2 pts) Compute $I(X; Y)$. (You may express your answer in terms of $\log_2$ of an appropriate integer.)

   We have $I(X; Y) = H(Y) - H(Y|X)$. We have $p(Y = 1) = 1/4 + 0 = 1/4$. Also, $p(Y = 2) = 1/4 + 1/2 = 3/4$. So, $I(X; Y) = H_2(1/4) - 0.5$.

2. (6 points) Let X_1 and X_2 be random variables with possible outcomes $\mathcal{X}$. Suppose that these variables are identically distributed, i.e. that $P(X_1 = x) = P(X_2 = x)$ for all $x \in \mathcal{X}$. However, we do *not* assume X_1 and X_2 are independent. Now let

   (a) (2 pts) Prove t $\frac{\quad}{H(X_1)}$

   This is because $H(X_1) - H(X_2|X = 1) =$ $= 1) = I(X_1; X_2)$.

   (b) (2 pts) Hence, or otherwise, prove that 0

   This is because $0 \leq I(X_1; X_2) \leq H(X_1)$.

   (c) (1 pt) When is $\rho = 0$? Justify your answer.

   This is when $I(X_1; X_2) = 0$, i.e. when the two variables are independent.

   (d) (1 pt) When is $\rho = 1$? Justify your answer.

   This is when $I(X_1; X_2) = H(X_1)$, or $H(X_2|X_1) = H(X_2)$, i.e. when the two variables are identical.

3. (6 points) The post office in Union Court of ANU handles 10,000 letters per day on average.

   (a) (3 pts) Use Markov's inequality to derive an upper bound on the probability that more than 12,000 letters will be handled tomorrow. (Leave your answer as a fraction.)

   This is $P(X \geq 12000) \leq \frac{10000}{12000} = \frac{5}{6}$.

   (b) (3 pts) Suppose the variance of letters processed is 2,000. Use Chebyshev's inequality to derive an upper bound on the probability that this post office will handle between 8,000 and 12,000 letters tomorrow. (Leave your answer as a fraction.)

   We know that

$$P(|X - 10000| \geq 2000) \leq \frac{2000}{4000000}.$$

The question then requires we compute $P(8000 < X < 12000) = P(|X - 10000| < 2000) = 1 - P(|X - 10000| \geq 2000)$.

4. (8 points) Recall that a 3-tuple of random variables (X, Y, Z) form a Markov chain if and only if

$$p(X, Y, Z) = p(X)\, p(Y \mid X)\, p(Z \mid Y).$$

Similarly, a 4-tuple of random variables (X, Y, Z, W) form a Markov chain if and only if

$$p(X, Y, Z, W) = p(X)\, p(Y \mid X)\, p(Z \mid Y)\, p(W \mid Z).$$

In what follows, suppose (X, Y, Z, W) forms a Markov chain.

(a) (2 pts) Prove that (X, Y, Z) forms a Markov chain.

We have

$$\begin{aligned} p(X, Y, Z) &= \sum_w p(X, Y, Z, W = w) \\ &= \sum_w p(X)\, p(Y \mid X)\, p(Z \mid Y)\, p(W = w \mid Z) \\ &= p(X)\, p(Y \mid X)\, p(Z \mid Y) \sum_w p(W = w \mid Z) \end{aligned}$$

This is p

in.

(b) (4 pts) Prove t

Note that since (X, Y, Z) is a Markov chain,). Keeping this in mind, we have

$$\begin{aligned} p(X, Z, W) &= \sum_y p(X, Y = y, Z, W) \\ &= \sum_y p(X)\, p(Y = y \mid X)\, p(Z \mid Y = y)\, p(W \mid Z) \\ &= p(X)\, p(W \mid Z) \sum_y p(Y = y \mid X)\, p(Z \mid Y = y) \\ &= p(X)\, p(W \mid Z) \sum_y p(Y = y \mid X)\, p(Z \mid Y = y, X) \\ &= p(X)\, p(W \mid Z) \sum_y p(Z, Y = y \mid X) \\ &= p(X)\, p(W \mid Z)\, p(Z \mid X). \end{aligned}$$

This is precisely the definition of (X, Z, W) being a Markov chain.

(c) (2 pts) Prove that $I(X; W) \leq I(Y; Z)$. (You may use without proof the following result from Tutorial 5: if (X, Y, Z) forms a Markov chain, then (Z, Y, X) also forms a Markov chain, and so $I(X; Z) \leq \min(I(X; Y), I(Y; Z))$. You may also use the results of (a) and (b), even if you are unable to prove them.)

From (a), (X, Y, Z) is a Markov chain. From the statement in tutorials, $I(X; Z) \leq I(Y; Z)$.

From (b), (X, Y, W) is a Markov chain. From the statement in tutorials, $I(X;W) \leq I(X;Z)$.
Combining the two inequalities gives the result.

SECTION C.

Answer each of the following questions [Marks per questions as shown; 25% total]

1. (5 pts) Suppose X is an ensemble over three possible outcomes, with probabilities $p_X = (0.4, 0.3, 0.3)$. Recall that X^N denotes an extended ensemble.

   (a) (1 pt) Compute the raw bit content $H_0(X)$. (You may express your answer in terms of $\log_2$ of an appropriate integer.)

   This is $\log_2 3$.

   (b) (1 pt) What quantity, if any, does $\frac{1}{N} H_0(X^N)$ converge to as $N \to \infty$? Justify your answer.

   We have $H_0(X^N) = \log |A_{\mathcal{X}^N}| = \log |A_{\mathcal{X}}^N| = N \log |A_{\mathcal{X}}| = NH_0(X)$. So this stays at the constant value $H_0(X)$.

   (c) (2 pts) Compute the essential bit content $H_\delta(X)$ when $\delta = 0.35$.

   If $\delta = 0.35$ we want the smallest set with at least 0.65 of the probability mass. This is obtained by throwing out either element with smallest probability 0.3, leaving behind two elements. We thus find $H_\delta(X) = \log_2 2 = 1$.

   (d) (1 pt) What quantity, if any, does $\frac{1}{N} H_\delta(X^N)$ converge to as $N \to \infty$ for $\delta = 0.35$? Justify your answer.

   The entropy $H(X)$, by the source coding theorem.

2. (12 pts) Suppose bilities $p_X = (0.3, 0.3, 0.2,$

   (a) (4 pts) Comp g.

   Initially we merge d and e to get a new symbol with pr "de" and c to get a new symbol with probability 0.4. T a new symbol with probability 0.6. Then we merge t $\to$ 10, b $\to$ 11, c $\to$ 01, d $\to$ 000, e $\to$ 001.

   (b) (2 pts) Compute the expected length $L(C, X)$ for your Huffman code.

   This is $2 \cdot 2 \cdot 0.3 + 2 \cdot 0.2 + 2 \cdot 3 \cdot 0.1 = 1.2 + 0.4 + 0.6 = 2.2$.

   (c) (2 pts) Explain the relationship between $L(C, X)$ and the entropy $H(X)$.

   We know $H(X) \leq L(C, X)$, so $H(X) \leq 2.2$. In fact $H(X) = 2.1710$.

   (d) (1 pt) Guffman, a self-trained mathematician, claims he has discovered a new prefix code C' which has expected length $L(C', X)$ strictly smaller than $L(C, X)$. By referencing an appropriate theorem from lectures, explain whether his claim is possible.

   Not possible. Huffman codes have shortest expected length out of all prefix codes.

   (e) (1 pt) Hoffman, a self-trained quant, claims that he has constructed a new prefix code C'' has codeword lengths $(1, 1, 2, 2, 2)$. By referencing an appropriate theorem from lectures, explain whether his claim is possible.

   Not possible, because prefix code implies $\sum 2^- \ell \leq 1$, not the case here.

   (f) (2 pts) Suppose we compute a Shannon code C''' for X. Should we expect $L(C''', X) = H(X)$? Explain why or why not.

   No, because the probabilities are not powers of two, so the codeword lengths will be larger than the log probabilities.

3. (6 pts) Suppose X is an ensemble over outcomes $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$. Let the probabilities $p_X = (0.25, 0.5, 0.125, 0.125)$.

   (a) (2 pts) Compute the codeword lengths for all outcomes under a Shannon-Fano-Elias code for X.

   Lengths are $\lceil \log 1/p(x) \rceil + 1 = (3, 2, 4, 4)$.

   (b) (4 pts) Compute the Shannon-Fano-Elias codewords for $\mathtt{a}$ and $\mathtt{b}$, showing all your working. (You do not need to compute codewords for $\mathtt{c}$ and $\mathtt{d}$.)

   We have $\bar{F}(a) = 0.125 = 0.001$, $\bar{F}(b) = 0.25 + 0.25 = 0.5 = 0.10$. So, a $\rightarrow$ 001, b $\rightarrow$ 10.

4. (2 pts) Briefly describe one potential advantage of arithmetic coding over Huffman coding.

   It can adapt to changing probability distributions, and does not assume the probabilities stay the same at every single iteration.

SECTION D.

Answer each of the following questions [Marks per questions as shown; 25% total]

1. (2 pts) Consider a binary symmetric channel with bit flip probability $f = 0.25$.

   (a) (1 pt) Write down the transition matrix Q for this channel.

   $Q = \begin{bmatrix} 1-f & f \\ f & 1-f \end{bmatrix} = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$.

   (b) (1 pt) What input distribution achieves the capacity of the channel? (You may refer to a result from lectures.)

   A uniform distribution, since the channel is symmetric.

2. (5 pts) Let Q denote some channel over input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, 1\}$.

   (a) (1 pt) Donald computes the mutual information $I(X; Y)$ using an input distribution $p_{\mathcal{X}} = (0.5, 0.5)$. He finds that for this distribution, $I(X; Y) = 0.9$. Based on this fact, what is a lower bound on the capacity of Q? Justify your answer.

   The capacity is the maximal $I(X; Y)$. So, the capacity is at least 0.9.

   (b) (2 pts) Provide an upper bound on the capacity of Q. Justify your answer.

   The capacity is the maximal $I(X;Y)$, which is at most $H(Y)$, which is at most 1 bit.

   (c) (2 pts) Carly c 1.8 bits per transmis m possible? Explain

   No, by the NCCT we cannot achieve rates above the c at most 1 bit per transmission.

3. (14 pts) Consider a channel over inputs $\mathcal{X} =$ $\mathcal{Y} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, with transition matrix

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

   and input distribution $p_{\mathcal{X}} = (p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}}, p_{\mathtt{d}})$.

   (a) (2 pts) Is the channel Q symmetric? Explain why or why not.

   No. We cannot partition the outputs such that the rows and column are permutations of each other. e.g. if we try the first two outputs, then the rows are permutations of each other, but not the columns.

   (b) (3 pts) Compute the probability distribution $p(Y)$ over outputs. (Express your answer in terms of $p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}}, p_{\mathtt{d}}$.)

   First, $p(Y = a) = p(Y = b) = (p_a + p_b)/2$. Next, $p(Y = c) = p_c$, and $p(Y = d) = p_d$.

   (c) (3 pts) Compute the conditional entropy $H(Y \mid X)$. (Express your answer in terms of $p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}}, p_{\mathtt{d}}$.)

   First, $H(Y|X = a) = H(Y|X = b) = H(1/2) = 1$. Next, $H(Y|X = c) = H(Y|X = d) = 0$. So, $H(Y|X) = p_a + p_b$.

(d) (3 pts) Hence, show that

$$I(X;Y) = -(1 - p_{\mathtt{c}} - p_{\mathtt{d}}) \cdot \log_2(1 - p_{\mathtt{c}} - p_{\mathtt{d}}) - p_{\mathtt{c}} \cdot \log_2 p_{\mathtt{c}} - p_{\mathtt{d}} \cdot \log_2 p_{\mathtt{d}}.$$

We know $I(X;Y) = H(Y) - H(Y|X)$. First,

$$H(Y) = -(p_a + p_b) \cdot \log(p_a + p_b) + (p_a + p_b) - p_c \log p_c - p_d \log p_d.$$

So,

$$I(X;Y) = -(p_a + p_b) \cdot \log(p_a + p_b) - p_c \log p_c - p_d \log p_d.$$

Now clearly $p_a + p_b = 1 - p_c - p_d$. So, the result follows.

(e) (3 pts) What input distribution achieves the capacity of Q? Explain your answer intuitively.

The above is just the standard entropy for a random variable with distribution (p_c, p_d, p_e) where $p_e = 1 - p_c - p_d$. So, it must be maximised by a uniform distribution over these outcomes, i.e. with $p_c = p_d = p_e = 1/3$. Note here that the precise choice of p_a, p_b does not matter: by picking $p_c = p_d = 1/3$ we constrain the sum of these two outcomes to be $1/3$, but otherwise the precise values are arbitrary. This makes sense, because outcomes `a` and `b` are essentially interchangeable, and can be considered as one new symbol.

4. (4 pts) Suppose we use a (7, 4) Hamming code to communicate over a binary symmetric channel with no

(a) (3 pts) Comp ou may use a diagram to show your working.

It should be 110.

(b) (1 pt) Suppose a receiver sees the bit string $\mathtt{b_1 b_2 b_3}$ is the parity bit string you computed above. Is it guaranteed that the sender actually transmitted `1001`? Explain why or why not.

No, it is not guaranteed. There could have been three or more bit flips starting from another codeword.