SECTION A.

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

1. (12 points) Suppose X and Y are random variables with the following joint distribution:

$X \backslash Y$	1	2
0	0.2	0.1
1	0.0	0.2
2	0.3	0.2

Compute each of the following, showing all your working.

(a) (2 points) Compute the marginal distributions of X and Y, i.e. determine the values of $\mathbb{P}(X = x)$ for all x, and $\mathbb{P}(Y = y)$ for all y.

$x = 0, 1, 2, \mathbb{P}(X = x) = 0.3, 0.2, 0.5$

(b) (2 points) Compute the conditional distribution of X given that the value $Y = 2$, i.e. compute the values of $\mathbb{P}(X = x|Y = 2)$ for all x.

$x = 0, 1, 2, \mathbb{P}(X = x|Y = 2) = 0.2, 0.4, 0.4$

(c) (2 points) Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

$\mathbb{E}[X] =$ ... $] = 1 \cdot 0.5 + 2 \cdot 0.5 = 1.5$

(d) (4 points) Co...

$\mathbb{E}[XY]$ ... $\times 2) \times 0.2 + (2 \times 1) \times 0.3 + (2 \times 2) \times 0.2 = 1.8$

(e) (2 points) Are X and Y independent? Expl...

No, they are not independent, as can be seen from the f... $\mathbb{P}(X = 1, Y = 1) \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 1)$. Left side is 0 but rhs is $0.2 \times 0.5 = 0.1$

2. (5 points) You have a large bucket which contains 999 fair coins and one biased coin. The biased coin is a two-headed coin (i.e. it always lands heads). Suppose you pick one coin out of the bucket, flip it 10 times, and get all heads.

(a) (1 point) State Bayes' rule.

$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$ where A and B are events and $\mathbb{P}(B) \neq= 0$

(b) (1 point) State the Law of Total Probability.

$\mathbb{P}(X = x) = \sum_j \mathbb{P}(X = x_i, Y = y_j)$

(c) (3 points) What is the probability that the coin you choose is the two-headed coin?

Let the event that it is a double headed coin be B where B is the short-hand for biased. We want to compute $\mathbb{P}(B|10H)$. So, by Bayes rule and Law of Total Probability

$$\begin{aligned}\mathbb{P}(B|10H) &= \frac{\mathbb{P}(10H|B)\mathbb{P}(B)}{\mathbb{P}(10H)} \\ &= \frac{1 \cdot \frac{1}{1000}}{\mathbb{P}(10H|B)\mathbb{P}(B) + \mathbb{P}(10H|B^c)\mathbb{P}(B^c)}\end{aligned}$$

$$= \frac{1 \cdot \frac{1}{1000}}{1 \cdot \frac{1}{1000} + \left(\frac{1}{2}\right)^{10} \frac{999}{1000}}$$

$$\approx \frac{1}{2}.$$

3. (5 points)

   (a) (1 point) In one or two sentences, explain what is the difference between the Bayesian and Frequentist approaches to parameter estimation.

   Frequentist: parameters are fixed (but unknown); there is no distribution over them
   Bayesian: parameters are random, and have a corresponding distribution.

   (b) (1 point) In one or two sentences, explain what the maximum likelihood estimate for a parameter is.

   The maximum likelihood estimator (MLE),

$$\hat{\theta}(x) = \arg\max_{\theta} L(\theta|\mathbf{x})$$

   (c) (1 point) In one or two sentences, explain what is a prior distribution in the context of Bayesian inference.

   An uncert ... s one's beliefs ab ... ion.

   (d) (1 point) In on ...) estimate for a parameter is.

   We maximize the posterior or log posterior

   (e) (1 point) In one or two sentences, explain the connectio ... MAP estimates. (*Hint*: what priors should you use?)

   Uniform distribution

4. (3 points) A biased coin is flipped until the first head occurs. Denote by Z the number of flips required with p being the probabilty of obtaining a head. Compute $\mathbb{P}(Z = k)$ stating clearly the possible values of k.

   (*Hint*: Write down the probabilities of some possible outcomes.)

   when x = 1, success. when x = 2, fail then success, when x= 3, fail fail success

$$\mathbb{P}(Z = k) = q^{k-1} p$$

   for $k = 1, 2, 3, \ldots$

SECTION B.

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

5. (4 points) Suppose the random variables X and Y are related by the following probabilities $\mathbb{P}(X = x, Y = y)$:

X \ Y	0	1
0	1/3	1/3
1	0	1/3

Compute the following:

(a) (2 points) $H(X)$, $H(Y)$.

(b) (2 points) $H(X|Y)$, $H(Y|X)$.

First compute the marginal distributions: $\mathbb{P}(x) = (2/3, 1/3)$ and $\mathbb{P}(y) = (1/3, 2/3)$, we now have

(a)

8 bits.

also simi

$H(Y) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) -$ ... s.

(b) We have

$$H(X|Y) = \frac{1}{3} \cdot H(X|Y = 0) + \frac{2}{3} \cdot H(X|Y = 1) = \frac{1}{3}H(1, 0) + \frac{2}{3}H(1/2, 1/2) = 2/3$$

similarly for $H(Y|X)$, we get

$$H(Y|X) = \frac{2}{3} \cdot H(Y|X = 0) + \frac{1}{3} \cdot H(Y|X = 1) = \frac{2}{3}H(1/2, 1/2) + \frac{1}{3}H(0, 1) = 2/3$$

6. (5 points)

(a) (3 points) What is a typical set? How does the typical set relate to the smallest delta-sufficient subset?

The typical set is a set of sequences whose probability is close to two raised to the negative power of the entropy of a random variable of interest.

Relationship: it is used in the proof of SCT, as $n \to \infty$, S_δ and typical set increasingly overlap. Hence, we look to encode all typical sequences uniformly, and relate that to the essential bit content by taking the log of smallest delta-sufficient subset.

(b) (2 points) Is the most likely sequence always a member of the typical set? Explain your answer.

The most likely sequence is in general not in the typical set. For example for X_k iid with $\mathbb{P}(0) = 0.1$ and $\mathbb{P}(1) = 0.9$, (1,1,1,...,1) is the most likely sequence, but it is not typical because its empirical entropy is not close to the true entropy.

7. (3 points) Construct a Huffman code for the ensemble with alphabet $\mathcal{A}_X = \{\mathsf{a}, \mathsf{b}, \mathsf{c}\}$ and probabilities $\mathbf{p} = (0.6, 0.3, 0.1)$. Show all your working.

The Huffman code for the distribution is $(0.6, 0.3, 0.1)$ is $(1, 01, 00)$

8. (7 points) Suppose a single fair six-sided die is rolled, and let Y be the outcome.

(a) (3 points) Compute the expected value of Y and the variance of Y.

Use standard formula: $\mathbb{E}(Y) = \frac{1}{6}(1+2+3+4+5+6) = 3.5$ and $\mathbb{V}ar(Y) = \frac{1}{6}\left[(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right] = \frac{35}{12}$.

(b) (1 point) Calculate an upper bound for the quantity $\mathbb{P}(Y \geq 6)$ using Markov's inequality.

By Markov's inequality, we get:

$$\mathbb{P}(Y \geq 6) \leq \frac{3.5}{6} = 0.583.$$

(c) (3 points) Ca... 6) using Chebyshev's inequali... (b). Which is closer to the true value of $\mathbb{P}(Y \geq 6)$?

By Chebyshev's inequality, we get:

$$\mathbb{P}(Y \geq 6) \leq \mathbb{P}(Y \geq 6 \text{ or } Y \leq 1) = \mathbb{P}(|Y - 3.5| \geq 2.5) \leq \frac{3.5}{2.5^2} = \frac{7}{15} = 0.467.$$

Chebyshev's is closer.

9. (6 points)

(a) (1 point) What is the purpose of the sigmoid function in logistic regression?

To squash the score to be in the range of 0 to 1

(b) (2 points) Suppose you have a trained logistic regression model with weights $\mathbf{w}$. Roman proposes to classify a new point $\mathbf{x}_{new}$ as positive by checking if $\mathbf{x}_{new}^T\mathbf{w} > 0$. Alice proposes to classify it as positive by checking if $\sigma(\mathbf{x}_{new}\mathbf{w}) > 0.5$, where $\sigma(\cdot)$is the sigmoid function. Will these result in the same prediction? Explain why or why not.

Yes they will result in the same prediction. This is because the sigmoid is a monotone increasing function, and $sigmoid(0) = 0.5$. Hence $sigmoid(x) > 0.5$ iff $x > 0$.

(c) (3 points) In one or two sentences, explain the relationship between logistic regression and the maximum entropy principle.

(1) The maxent principle for estimating probabilities: 'When choosing amongst multiple possible distributions, pick the one with highest entropy'. (2) Given information

about a probability distribution, we can find the maximum entropy distribution using Lagrangian optimisation. (3) Logistic regression can be derived from the (conditional) maximum entropy principle

SECTION C.

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

10. (10 points) Suppose Y is an ensemble equipped with $\mathcal{A}_Y = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and probabilities $\mathbf{p} = (0.5, 0.25, 0.25)$.

    (a) (2 points) Write down the alphabet and probabilities for the extended ensemble Y^2.

    We have that

    $$\mathcal{A}_Z = \{aa, ab, ba, ac, ca, bc, cb, bb, cc\}$$
    $$\mathbf{p} = \{0.25, 0.125, 0.125, 0.125, 0.125, 1/16, 1/16, 1/16, 1/16\}$$

    (b) (3 points) Assuming the symbols for Y^2 are in alphabetical order, so that e.g. `aa` appears before `ab`, what is the binary interval for `ab` in a Shannon-Fano-Elias code for Y^2?

    Alphabetical order means that ab will be the second, so that cdf gives

    $$F(aa) = 0.25, F(ab) = 0.375$$

    so the interval is

    $$[0.25, 0.374) - \text{decimal}$$

    (c) (3 points) What is the smallest δ−suffici 2 = 0.45?

    $\mathbb{P}(Y \in S_\delta) \geq 1 - 0.45 = 0.55$ so $S_\delta = \{a$

    (d) (2 points) What is the essential bit content 5?

    $H_\delta(Y^2) = \log_2 |S_\delta| = 2.$

11. (6 points)

    (a) (1 point) Is every prefix-free code uniquely decodable? If yes, explain why. If no, provide a counter-example.

    A prefix code is uniquely decodable, i.e. given a complete and accurate sequence, a receiver can identify each word without requiring a special 'marker' between the words.

    This is an inductive argument: suppose the messages x and y disagree at the Kth symbol. Then by definition the codewords for the Kth symbol must disagree, since no codeword can be a prefix of another. Hence the codes of the entire message must be different.

    (b) (1 point) Is the code $C = \{0, 01, 011\}$ uniquely decodable? Explain your answer.

    Yes - as we move along the message, we uncover the first, second and third codeword.

    (c) (2 points) Explain the difference between a lossless and uniquely decodable code.

    Recall that a code is lossless if for all $x, y \in \mathcal{A}_X$

    $$x \neq y \implies c(x) \neq c(y)$$

This ensures that if we work with a single outcome, we can uniquely decode the outcome. When working with variable-length codes, however, unique decodability is defined as follows: A code c for X is **uniquely decodable** if no two strings from $\mathcal{A}_X$ have the same codeword. That is, for all $\vec{x}, \vec{y} \in \mathcal{A}_X$

$$\vec{x} \neq \vec{y} \implies c(\vec{x}) \neq c(\vec{y})$$

The crux of the matter: one is a number x but the other is a vector $\vec{x}$.

(d) (2 points) Consider a source $W = \{a, b, c, d, e\}$. Explain if it is possible to construct a prefix code for this source with the proposed lengths: $l_a = 1$, $l_b = 2$, $l_c = 3$, $l_d = 4$, $l_e = 4$, without actually giving an example of a code? (*Hint*: What conditions should you check?)

It satisfies the Kraft's inequality with exact equality, $\sum_x 2^{-l_x} = \frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^4} = 1$, so yes we can construct a prefix code for this source.

12. (9 points) Let X be an ensemble with alphabet $\mathcal{A}_X = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$ with probabilities $\mathbf{p} = (1/2, 1/4, 1/8, 1/8)$ and the code $C = (0000, 01, 11, 0)$.

(a) (2 points) What is the entropy $H(X)$ as a single number?

(b) (2 points) What is the expected code length $L(C, X)$?

(c) Which of these are Shannon codes? Justify your answers.

i. (1 poin

ii. (1 poin

iii. (1 poin =

(d) (2 points) Is the code A in part (c)[i] opti

(a) $H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 = 1.75$

(b) $L(C, X) = \frac{1}{2} \times 4 + \frac{1}{4} \times 2 + \frac{1}{8} \times 2 + \frac{1}{8} \times 1 = 2.875 = 2\frac{7}{8}$

(c) Shannon codes for X has the following code length:

i. (1 point) $A = \{0, 10, 110, 111\}$ - yes

ii. (1 point) $B = \{000, 001, 010, 111\}$ - no

iii. (1 point) $C = \{0, 01, 001, 010\}$ - no

(d) It is optimal because $L(A, X) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + 2 \times \frac{1}{8} \times 3 = 1.75 = H(X)$. By SCT this is optimal

SECTION D.

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

13. [5 points] Consider a $(5, 3)$ block code C.

    (a) (3 points) What is the length of each codeword in class C? How many codewords does class C define? Compute the rate for class C.

    There are $2^3 = 8$ codewords, each with length 5. The rate is $3/5 = 0.6$.

    (b) (2 points) Do there exist codes with rate equal to that of class C, that can achieve arbitrarily small probability of block error over a channel of capacity 0.4? Justify your answer.

    No, there cannot exist such codes, as the rate would exceed the channel capacity and violate the channel coding theorem.

14. [15 points] Let $\mathcal{X} = \{\mathsf{a}, \mathsf{b}\}$ and $\mathcal{Y} = \{\mathsf{a}, \mathsf{b}\}$ be the input and output alphabets for the following two channels:

$$R = \begin{bmatrix} 1 & 0.5 \\ 0 & 0.5 \end{bmatrix} \quad \text{and} \quad R^\dagger = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 1 \end{bmatrix}$$

(a) (2 points) Describe the behaviour of R and $R^\dagger$ when each of the symbols a and b are used as input.

R trans .5. The opposite for $R^\dagger$.

(b) Define an arb and $p + q = 1$ over $\mathcal{X}$. Express t

i. (2 points) The probabilities $\mathbb{P}(Y = \mathsf{a}$ f p, where Y is a random variable denoting the output of the c

$P(y = a) = \frac{1}{2}(1 + p)$ and $P(y = b)$ –

ii. (2 points) The entropy of $H(Y)$ in terms of the probability p and the function $H_2(\cdot)$ defined by

$$H_2(\vartheta) = -\vartheta \cdot \log_2 \vartheta - (1 - \vartheta) \cdot \log_2(1 - \vartheta).$$

$H(Y) = H_2(\frac{1}{2}(1 + p))$.

iii. (2 points) The mutual information $I(X; Y)$ in terms of p and the function H_2 defined above.

$I(X; Y) = H(Y) - H(Y|X) = H_2(\frac{1}{2}(1 + p)) - (1 - p)$.

(c) (4 points) Using the previous results or otherwise, compute the input distribution that achieves the channel capacity for R.

First calculate the derivative w.r.t. ϑ :

$$H_2'(\theta) = -\log \frac{\theta}{1 - \theta}.$$

From previous question we have $I(X; Y)$ as a function of p. As $I(X, Y)$ is a concave function of p. To maximise $I(X : Y)$, we solve

$$0 = \frac{d}{dp} I(X; Y)$$

$$= \frac{1}{2} \cdot H_2'(\frac{1}{2}(1+p)) + 1$$
$$= -\frac{1}{2} \cdot \log \frac{1+p}{1-p} + 1,$$

and so we need

$$\frac{1+p}{1-p} = 4$$

yielding $p = 0.6$. This gives $C(Q) \approx 0.32$.

(d) (3 points) Suppose you used the channels R and $R^\dagger$ to send messges by first flipping a fair coin and sending a symbol through R if it landed heads and through $R^\dagger$ if it landed tails. Construct a matrix Q that represents the channel defined by this process.

$$Q = \frac{1}{2}R + \frac{1}{2}R^\dagger = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

15. (5 marks) For an arbitrary noisy channel Q with N input symbols and M output symbols, show that its capacity ρ satisfies

$$\rho \leq \min\{\log_2 N, \log_2 M\}.$$

$$_2 \quad = \log_2 N$$

Similarly

$$\rho = I(X;Y) = H(Y) - H(Y|X) \leq \quad \leq \quad _2| \quad | = \quad g_2 M$$

so, we have

$$\rho \leq \min\{\log_2 N, \log_2 M\}.$$

– – – END OF EXAM – – –