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.

   (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.

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

   (d) (4 points) Compute $\mathbb{E}[XY]$.

   (e) (2 points) Are X and Y independent? Explain your answer.

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.

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

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

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.

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

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

   (d) (1 point) In one or two sentences, explain what the Maximum a posteriori (MAP) estimate for a parameter is.

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

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.)

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 $\mathbb{P}(X = x, Y = y)$:

$X \backslash Y$	0	1
0	1/3	1/3
1	0	1/3

Compute the following:

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

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

6. (5 points)

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

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

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.

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.

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

(c) (3 points) Calculate an upper bound for the quantity $\mathbb{P}(Y \geq 6)$ using Chebyshev's inequality and compare this to the answer obtained in part (b). Which is closer to the true value of $\mathbb{P}(Y \geq 6)$?

9. (6 points)

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

(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.

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

    (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?

    (c) (3 points) What is the smallest $\delta-$sufficient subset for Y^2 when $\delta = 0.45$?

    (d) (2 points) What is the essential bit content $H_\delta(Y^2)$ when $\delta = 0.45$?

11. (6 points)

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

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

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

    (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?)

12. (9 points) Let X be an ensemble with alphabet $\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$ 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 point) $A = \{0, 10, 110, 111\}$

       ii. (1 point) $B = \{000, 001, 010, 111\}$

       iii. (1 point) $C = \{0, 01, 001, 010\}$

    (d) (2 points) Is the code A in part (c)[i] optimal? Explain why or why not.

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.

    (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.

14. [15 points] Let $\mathcal{X} = \{\mathtt{a}, \mathtt{b}\}$ and $\mathcal{Y} = \{\mathtt{a}, \mathtt{b}\}$ 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 $\mathtt{a}$ and $\mathtt{b}$ are used as input.

    (b) Define an arbitrary distribution $\mathbf{p}_{\mathcal{X}} = (p, q)$ with $p \in [0, 1]$ and $p + q = 1$ over $\mathcal{X}$. Express the following for R:

        i. (2 points) The probabilities $\mathbb{P}(Y = \mathtt{a})$ and $\mathbb{P}(Y = \mathtt{b})$ in terms of p, where Y is a random variable denoting the output of the channel $\mathcal{Y}$.

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

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

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

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

    (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.

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

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

– – – END OF EXAM – – –