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

1. (8 pts) A scientist conducts 1000 trials of an experiment involves variables X, Y, each with possible values $\{0, 1\}$. He records the outcomes of the trials in the following table.

Counts		X	
		0	1
Y	0	50	400
	1	500	50

Compute each of the following, showing all your working.

(a) (1 pt) $p(X = 1, Y = 1)$.
By definition, $p(X = 1, Y = 1) = \frac{50}{1000} = 0.05$.

(b) (2 pt) $p(X = 1)$.
By the sum rule, $p(X = 1) = p(X = 1, Y = 1) + p(X = 1, Y = 0) = \frac{450}{1000} = 0.45$.

(c) (1 pt) $E[X]$.
By definition of expectation, $E[X] = p(X = 1) \cdot 1 + p(X = 0) \cdot 0 = p(X = 1) = 0.45$.

(d) (2 pt) $p(Y = 1|X = 1)$.
By definition of conditional probability, $p(Y = 1|X = 1) = \frac{p(X=1,Y=1)}{p(X=1)} = \frac{50}{450} = \frac{1}{9}$.

(e) (2 pt) $p(X + Y = 1)$.
We have $p(X + Y = 1) = p(X = 0, Y = 1) + p(X = 1, Y = 0) = \frac{900}{1000} = 0.9$.

2. (9 pts) Lorelai is taking an exam that involves multiple choice questions, with available choices $\{1, 2, \ldots, c\}$, where $c \geq 2$. For each question, there is only one correct answer, and Lorelai is allowed to provide only one choice as her answer.

With 90% probability, Lorelai knows the answer for a question. When Lorelai knows the answer for a question, she always picks the correct choice. When Lorelai does not know the answer for a question, she makes a guess. When making a guess, Lorelai is equally likely to pick any of the c available choices.

(a) (2 pt) Show that the probability Lorelai does not know the answer for a question and guesses the correct answer is

$$\frac{0.1}{c}.$$

Let $A \in \{0, 1\}$ denote whether the question is answered correctly, and $K \in \{0, 1\}$ denote whether the student knew the answer. From the information provided,

$$p(A = 1|K = 0) = \frac{1}{c}$$
$$p(K = 1) = 0.9.$$

Thus,

$$p(A = 1, K = 0) = p(A = 1|K = 0)p(K = 0) = \frac{1}{c} \cdot (1 - 0.9) = \frac{0.1}{c}.$$

(b) (2 pt) Show that the probability Lorelai answers a question correctly is

$$0.9 + \frac{0.1}{c}.$$

We have

$$\begin{aligned} p(A=1) &= p(A=1|K=1)p(K=1) + p(A=1|K=0)p(K=0) \\ &= 1 \cdot 0.9 + \frac{1}{c} \cdot (1 - 0.9) \\ &= 0.9 + \frac{1}{c} \cdot 0.1. \end{aligned}$$

(c) (3 pt) Suppose Lorelai answers a question correctly. Show that the probability she knew the answer is

$$\frac{9c}{9c+1}.$$

We are interested in $p(K=1|A=1)$. Using Bayes' rule, and part (b), we get

$$\begin{aligned} p(K=1|A=1) &= \frac{p(A=1|K=1)p(K=1)}{p(A=1)} \\ &= \frac{1 \cdot 0.9}{0.9 + c^{-1}(0.1)} \\ &= \frac{9}{9 + c^{-1}(1)} \\ &= \frac{9c}{9c+1}. \end{aligned}$$

(d) (2 pt) When c gets larger, does the probability in part (c) get larger or smaller? Explain your answer intuitively in one or two sentences.

It gets larger. When c is big, the more likely it is that Lorelai knows the answer, because it's unlikely she gets the answer by just guessing.

3. (8 pts)

(a) (2 pt) In one or two sentences, explain what the likelihood function for a parameter is.

The likelihood function is the density function regarded as a function of $\theta \in \Theta$, the set of parameters to be estimated. Concretely, the sample $X = (X_1, X_2, ..., X_n)$ is the sample X of random variables chosen according to one of a family of probabilities P_θ. The likelihood function is the is a function of θ:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta).$$

(b) (2 pt) 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) (2 pt) Suppose we have a coin which comes up heads with unknown probability p. We perform 5 independent flips of the coin, and observe the outcomes

$$\{\texttt{heads}, \texttt{heads}, \texttt{tails}, \texttt{heads}, \texttt{heads}\}.$$

Based on this data, what is the maximum likelihood estimate for the probability p?

The MLE for n Bernoulli trials is $\hat{p} = n^{-1} \sum_{i=1}^{n} X_i$, that is the average fraction of successes. In this case we have $\hat{p} = 4/5$.

(d) (2 pt) Suppose our prior belief in the coin coming up heads follows a Beta(a, b) distribution for certain constants a, b. In one or two sentences, describe a procedure which can additionally incorporate this prior belief to estimate the probability p. (You do not need to compute the estimate for p using this procedure.)

We could use a Bayesian method, which would the posterior distribution of the parameter p, $\Pr(p \mid \text{Data}) \propto L(p \mid \text{Data}) \cdot \Pr(p)$. We could in particular compute the mode of the posterior distribution, giving the maximum a-posteriori estimate. In the case of a Bernoulli random variable, this corresponds to adding "fake" occurrences of heads and tails and then computing the fraction of heads as usual.

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

4. (9 pts) Suppose X represents whether it is sunny or cloudy in Canberra, and Y represents whether it is warm or humid in Canberra. The joint distribution $p(X, Y)$ is

$$\begin{aligned} p(X = \texttt{sunny}, Y = \texttt{warm}) &= 1/2 \\ p(X = \texttt{sunny}, Y = \texttt{humid}) &= 1/4 \\ p(X = \texttt{cloudy}, Y = \texttt{warm}) &= 1/4 \\ p(X = \texttt{cloudy}, Y = \texttt{humid}) &= 0. \end{aligned}$$

   (a) (2 pt) In one or two sentences, explain what the entropy of a random variable represents.
   It is the inherent uncertainty in the random variable. The closer to a uniform distribution, the higher the uncertainty.

   (b) (3 pt) Compute the entropy $H(X)$.
   We have $p(X = \texttt{sunny}) = 1/2 + 1/4 = 3/4$. So, the entropy is

$$\begin{aligned} H(X) &= H_2(3/4) \\ &= \frac{3}{4} \cdot \log_2 \frac{4}{3} + \frac{1}{4} \cdot \log_2 4 \\ &= \log_2 4 - \frac{3}{4} \log_2 3 \\ &= 2 - \frac{3}{4} \log_2 3. \end{aligned}$$

   (c) (2 pt) Compute the joint entropy $H(X, Y)$.

$$H(X,Y) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4} + 0\log 0\right) = 1.5.$$

   (d) (2 pt) Compute the conditional entropy $H(Y|X)$.
   We have

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) \\ &= 1.5 - 2 + 0.75 \cdot \log_2 3 \\ &= 0.75 \cdot \log_2 3 - 0.5. \end{aligned}$$

5. (12 pts) Let X and Y be independent random variables with possible outcomes $\{0, 1\}$, with

$$p(X = 1) = \frac{1}{2}$$

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

   Let $Z = X + Y$.

(a) (1 pt) Compute $I(X;Y)$.

As the variables are independent, $I(X;Y) = 0$.

(b) (4 pt) Show that

$$p(X = 1|Z = 0) = 0$$
$$p(X = 1|Z = 1) = \frac{1}{2}$$
$$p(X = 1|Z = 2) = 1.$$

Consider the following possibilities:

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	2

Therefore:

$$\begin{aligned} p(X = 1|Z = 0) &= 0 \\ p(X = 1|Z = 2) &= 1 \\ p(X = 1|Z = 1) &= \frac{p(X = 1, Z = 1)}{p(Z = 1)} \\ &= \frac{p(X = 1, Y = 0)}{p(X = 1, Y = 0) + p(X = 0, Y = 1)} \\ &= \frac{1}{2}. \end{aligned}$$

(c) (3 pt) Compute $H(X|Z)$.

$$\begin{aligned} H(X|Z) &= \sum_{z \in \{0,1,2\}} p(Z = z) \cdot H(X|Z = z) \\ &= p(Z = 1) \cdot H(X|Z = 1) \\ &= p(Z = 1) \cdot H_2(1/2) \\ &= (p(X = 0, Y = 1) + p(X = 1, Y = 0)) \cdot 1 \\ &= \frac{1}{2}. \end{aligned}$$

(d) (2 pt) Compute $I(X;Z)$.

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &= 1 - \frac{1}{2} \\ &= \frac{1}{2}. \end{aligned}$$

(e) (2 pt) Explain why parts (a) and (d) do not contradict the data-processing inequality.
It doesn't because (X, Y, Z) doesn't form a Markov chain.

6. (4 pts) Dr. Seuss has set an exam for his class, where the possible scores are $\{0, 1, 2, \ldots, 100\}$ i.e. the exam is out of 100 marks. Dr. Seuss found that the average exam score of the students is 60 marks out of 100.

   (a) (2 pt) What is an upper bound on the fraction of students who scored 80 or more in the exam?
   Let X be the student score. Then $E[X] = 60$. Further, by Markov's inequality,

$$P(X \geq 80) \leq \frac{E[X]}{80} = \frac{60}{80} = \frac{3}{4}.$$

   (b) (2 pt) Dr. Preibus has set an exam for a different class, where the possible scores are $\{-100, -99, \ldots, 99, 100\}$ i.e. the exam is out of 100 marks, but it is possible to score negative marks. Dr. Preibus found that the average exam score of the students is 60 marks out of 100.
   Is your answer from (a) also an upper bound on the fraction of students who scored 80 or more in Dr. Preibus' exam? Explain why or why not.
   No, it isn't. Markov's inequality assumes the random variable is nonnegative.

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

7. (6 pt) Let X be an ensemble with $\mathcal{A}_X = \{\mathsf{a}, \mathsf{b}\}$ and probabilities $p_X = \left(\frac{1}{4}, \frac{3}{4}\right)$.

   (a) (1 pt) Write down the alphabet and probabilities for the extended ensemble X^2.
   We have $\mathcal{A}_{X^2} = \{\mathsf{aa}, \mathsf{ab}, \mathsf{ba}, \mathsf{bb}\}$ and $p_{X^2} = (1/16, 3/16, 3/16, 9/16)$.

   (b) (3 pt) Compute the essential bit content $H_\delta(X^2)$ for $\delta = \frac{3}{8}$.
   The extended ensemble X^2 has cumulative probabilities $(1/16, 1/4, 7/16, 1)$. We know that $\frac{1}{4} = \frac{4}{16} < \frac{3}{8} = \frac{6}{16} < \frac{7}{16}$. So, the essential bit content is $\log_2 2 = 1$.

   (c) (2 pt) Suppose $\delta = 0$. Is it true that $\frac{1}{N} H_\delta(X^N)$ will be arbitrarily close to the entropy $H(X)$ for N sufficiently large? Explain why or why not.
   No. $\frac{1}{N} H_0(X^N) = \frac{1}{N} \cdot \log |\mathcal{A}_{X^N}| = \log 2 = 1$. This falls outside the purview of the SCT, which requires $0 < \delta < 1$.

8. (10 pts) Let X be an ensemble with $\mathcal{A}_X = \{x_1, x_2, x_3\}$ and probabilities $p_X = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$.

   (a) (3 pt) Compute the modified distribution function $\bar{F}(x_i)$ for each outcome x_i.
   We have

$$\bar{F}(x_1) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} = 0.001_2$$
$$\bar{F}(x_2) = \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = 0.1_2$$
$$\bar{F}(x_3) = \frac{1}{4} + \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4} = \frac{7}{8} = 0.111_2.$$

   (b) (3 pt) Compute the Shannon-Fano-Elias codeword lengths $\ell(x_i)$ for each outcome x_i.
   We have

$$\ell(x_1) = \lceil \log_2 4 \rceil + 1 = 3$$
$$\ell(x_2) = \lceil \log_2 2 \rceil + 1 = 2$$
$$\ell(x_3) = \lceil \log_2 4 \rceil + 1 = 3.$$

   (c) (4 pt) Construct a Shannon-Fano-Elias code for X.
   We have

$$C(x_1) = 001$$
$$C(x_2) = 10$$
$$C(x_3) = 111.$$

9. (4 pts) Let X be an ensemble with $\mathcal{A}_X = \{\mathsf{a}, \mathsf{b}\}$ and fixed probabilities $p_X = (p_\mathsf{a}, p_\mathsf{b})$.

(a) (2 pt) Conan has constructed a prefix-free code C_1 for X, with expected length $L(C_1, X)$. Is it possible that $L(C_1, X) < H(X)$? Explain why or why not.

No, by the source coding theorem for symbol codes we cannot go below the entropy.

(b) (2 pt) Ronan has constructed a Huffman code C_2 for X. Is it possible that $L(C_2, X) > H(X)$? Explain why or why not.

Yes, Huffman codes don't have to exactly attain the entropy.

10. (5 pts) Suppose we wish to compress a string with LZ77 with a window size $W = 2$. Suppose we place two conditions on the string: it must have length 3 or more, and can only contain the symbols {a, b}. (For example, aba and bbb are valid strings, but ba (which has length 2) and abc (which contains the symbol c) are not.)

Recall that in LZ77, the output at any iteration is either of the type SYM (for an exact symbol match) or POINTER (for a window match).

(a) (2 pt) Give an example of a string satisfying the above conditions where, at every iteration of LZ77 except the first, the output is of the type POINTER.

Consider the string aaa.

(b) (3 pt) Is it possible that for some string satisfying the above conditions, at every iteration of LZ77, the output is of the type SYM? If yes, give an example of one such string. If no, explain why not.

No. Suppose we have SYM at the second iteration. Then the string starts with ab or ba. In either case, it is impossible for the third character to be anything else than a POINTER to something in the window.

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

11. (6 pts) Suppose we wish to construct a code using input alphabet $\mathcal{X} = \{0, 1\}$.

    (a) (1 pt) Give an example of a $(3, 2)$ block code using input alphabet $\mathcal{X}$.
    We need $2^2 = 4$ codewords of length 3. Pick e.g. $\{000, 001, 010, 011\}$.

    (b) (1 pt) What is the rate of your code from (a)?
    It is $\frac{K}{N} = \frac{2}{3}$.

    (c) (2 pt) Suppose you use your code from (a) on a channel with capacity 0.5. Is it possible that your code can achieve arbitrarily small probability of maximal block error? Explain why or why not.
    No. The NCCT says that we cannot achieve any rates about the capacity. And $\frac{2}{3} > 0.5$, so such a rate is non-achievable on this particular channel.

    (d) (2 pt) Suppose you use your code from (a) on a channel with unknown capacity. You find that the code can achieve arbitrarily small probability of maximal block error on this channel. What, if anything, can you conclude about the capacity of this channel?
    The NCCT says any achievable rate must be below the capacity. So, we know the capacity is at least $\frac{2}{3}$.

12. (14 pts) Consider a channel over inputs $\mathcal{X} = \{1, 2\}$ and outputs $\mathcal{Y} = \{1, 2\}$ with transition matrix

$$Q = \begin{bmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{bmatrix}.$$

    Let $p_X = (\theta, 1 - \theta)$ be a distribution over the inputs. Let $H_2(a)$ denote the entropy of a Bernoulli random variable with parameter a, i.e. $H_2(a) = -a \cdot \log a - (1 - a) \cdot \log(1 - a)$.

    (a) (2 pt) Is Q symmetric? Explain why or why not.
    No, we cannot partition into feasible rows.

    (b) (3 pt) Compute $p(Y = y)$ for $y \in \{1, 2\}$, expressing your answer in terms of θ.
    We can compute

$$\begin{aligned} p(Y = 1) &= \frac{3}{4}\theta + \frac{1}{2}(1 - \theta) \\ &= \frac{1}{4}\theta + \frac{1}{2} \\ p(Y = 2) &= \frac{1}{4}\theta + \frac{1}{2}(1 - \theta) \\ &= -\frac{1}{4}\theta + \frac{1}{2}. \end{aligned}$$

    (c) (2 pt) Compute $H(Y|X)$, expressing your answer in terms of θ.
    We have

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} H(Y|X = x) \cdot p(X = x) \\ &= \theta \cdot H_2(1/4) + (1 - \theta) \cdot H_2(1/2) \\ &= \theta \cdot H_2(1/4) + (1 - \theta). \end{aligned}$$

(d) (3 pt) Compute $I(X;Y)$, expressing your answer in terms of θ.
We have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2\left(\frac{1}{4}\theta + \frac{1}{2}\right) - (\theta \cdot H_2(1/4) + (1-\theta)) \\ &= H_2\left(\frac{1}{4}\theta + \frac{1}{2}\right) + \theta \cdot (1 - H_2(1/4)) - 1. \end{aligned}$$

(e) (3 pt) Show that the mutual information is maximised when θ satisfies

$$\frac{\frac{1}{2} + \frac{1}{4}\theta}{\frac{1}{2} - \frac{1}{4}\theta} = 2^{4\cdot(1-H_2(1/4))}.$$

Hint: Recall that $H_2(x)$ has derivative $H_2'(x) = -\log_2 \frac{x}{1-x}$.
We have

$$\frac{d}{d\theta} I(X;Y) = -\frac{1}{4} \cdot \log_2 \frac{\frac{1}{2} + \frac{1}{4}\theta}{\frac{1}{2} - \frac{1}{4}\theta} + (1 - H_2(1/4)),$$

so at optimality

$$\log_2 \frac{\frac{1}{2} + \frac{1}{4}\theta}{\frac{1}{2} - \frac{1}{4}\theta} = 4 \cdot (1 - H_2(1/4))$$

or

$$\frac{\frac{1}{2} + \frac{1}{4}\theta}{\frac{1}{2} - \frac{1}{4}\theta} = 2^{4\cdot(1-H_2(1/4))}.$$

(f) (1 pt) How do the answers to parts (d) and (e) relate to the channel capacity of Q?
The capacity is the mutual information in (d) at the value of θ in (e).

13. (3 pts) Calculate the value of the three parity bits for the message $x = 1111$ when it is coded using a $(7, 4)$ Hamming code. You may use a diagram to show your working.
The result should be 1111111.

14. (2 pts) Two distinct notions of uncertainty we have looked at are (prefix) Kolmogorov complexity K, and Shannon entropy H. What are the pros and cons of K versus H?
K is:

- not computable, unlike H.
- a key ingredient in Solomonoff induction, unlike H.

H is:

- dependent on a probability distribution, unlike K.
- a key ingredient in fundamental limits in coding and compression, unlike K.