

AUSTRALIAN NATIONAL UNIVERSITY

COMP2610/COMP6261

Information Theory, Semester 2 2023

Assignment 1

Due Date: Monday 28 August 2023, 9:05 am

Assignment 1 weighting is 10% of the course mark.

Instructions:

Marks:

1. The mark for each question is indicated next to the question. For questions where you are asked to prove results, if you can not prove a precedent part, you can still attempt subsequent parts of the question assuming the truth of the earlier part.
2. **COMP2610 students:** Answer *Questions 2, 3, 5* and only *Section I of Question 1 and 4*. You are not expected to answer Section II of Question 1 and 4. You will be marked out of 100.
3. **COMP6261 students:** Answer *Questions 2, 3, 5* and only *Section II of Question 1 and 4*. You are not expected to answer Section I of Question 1 and 4. You will be marked out of 100.

Submission:

1. Submit your assignment together **with a cover page** as a single PDF on Wattle.
2. **Clearly mention whether you are a COMP2610 student or COMP6261 student on the cover page.**
3. A late submission attracts a penalty of **5%** per working day as per ANU Policy until a **week (5 working days)** from the due date. We will provide solutions to the assignment after a week from the due date and if you submit after that time you get zero marks (**100% penalty**). Extensions will be considered according to the ANU Policy.
4. All assignments must be done individually. Plagiarism is a university offense and will be dealt with according to university procedures <http://academichonesty.anu.edu.au/UniPolicy.html>.

Question 1 [25 marks total]

****Section I: Only COMP 2610 students are expected to attempt this question.**

The Blue Challenge is a unique swimming competition that tests both the physical and emotional endurance of the participants and gives them an exciting opportunity to experience the oceanic wilderness.

Let's consider two universities, M (MIT) and C (Cambridge), competing in a series of 5 rounds. The contest rules are as follows:

- Each round has exactly one winner (i.e. universities cannot tie).
- In every race, both universities have equal probability ($\frac{1}{2}$) of winning.
- Once one university wins 3 rounds, no further rounds are conducted.

Let S be the random variable whose outcomes are all possible strings representing the winners of the rounds,

$$S = \{MMM, CCC, MCMM, \dots, MCMCM\}$$

Let T denote a random variable representing the total number of rounds conducted in the series, so that the possible outcomes are $\mathcal{T} = \{3, 4, 5\}$.

Calculate each of the following, showing all your working:

- (a) The information in seeing the sequence $MCMM$, i.e., compute $h(S = MCMM)$ [3 marks]
- (b) $H(S)$ [3 marks]
- (c) $H(T)$ [3 marks]
- (d) $H(S|T = 3)$ [3 marks]
- (e) $H(S|T)$ [4 marks]
- (f) $H(T|S)$. Provide an intuitive explanation of your answer. [3 marks]
- (g) Generally, the winner of the series is simply the university that wins the last round. However, if the university who wins the first round in the series is different to the one who wins the last round, then officials are suspicious of some form of fixing. In such a series, there is a 50% chance of officials declaring the series invalid, and neither university is declared winner; the other 50% of the time, the officials decide the series is legitimate, and declare the winner as usual.

Let P be a random variable denoting the identity of the winning university, with possible outcomes $\mathcal{U} = \{M, C, N\}$, for the winner being respectively MIT, Cambridge, or neither. Compute $H(U)$. [6 marks]

****Section II: Only COMP 6261 students are expected to attempt this question.**

In "The Trans-Galactic Olympiad", alien civilizations from different galaxies compete in a series of 5 rounds, testing both their physical and emotional resilience while giving them a thrilling opportunity to experience the interstellar wilderness.

Let's consider three civilizations, A (Andromeda), M (Milky Way), and S (Sombrero), participating in the tournament. The tournament rules are as follows:

- Each round has exactly one winner (i.e. civilizations cannot tie).
- In every round, all civilizations have an equal probability ($\frac{1}{3}$) of winning.
- Once a civilization wins 3 rounds, no further rounds are conducted.

Let S , be the random variable whose outcomes are all possible strings representing the winners of the rounds including when there is no final winner for 5 rounds.

Assumption: Now assume that by the end of 5 rounds, one civilization must emerge as the winner. Here, we are not accounting for outcomes where all three civilizations (A, M, S) failed to clinch 3 wins out of 5 rounds.

Let P be the random variable whose outcomes are all possible strings representing the winners of the rounds **as per the assumption**,

$$\mathcal{P} = \{\mathcal{AAA}, \mathcal{MMM}, \mathcal{SSS}, \dots, \mathcal{AMSA}\}$$

Note: Here, P is the subset of sample space S such that, $S = \{P, N \text{ (no winner)}\}$.

Let Q denote a random variable representing the total number of rounds conducted in the series, so that the possible outcomes are $Q = \{3, 4, 5\}$.

Calculate each of the following, showing all your working [Hint: You may need to use any programming tool (of your choice) to calculate the sample space \mathcal{P}]:

- The information in seeing the sequence $AMSSS$, i.e., compute $h(P = AMSSS)$ [2 marks]
- $H(P)$ [3 marks]
- $H(Q)$ [3 marks]
- $H(P|Q = 4)$ [3 marks]
- $H(P|Q)$ [3 marks]
- $H(Q|P)$. Provide an intuitive explanation of your answer. [3 marks]
- Generally, the winner of the series is simply the civilization that wins the last round. However, if the civilization who wins the first round in the series is different to the one who wins the last round, then the officials are suspicious of some form of competition manipulation. In such a series, there is a 33% chance of officials declaring the series

invalid, and no civilization is declared the winner; the other 66% of the time, the officials decide the series is legitimate, and declare the winner as usual.

Additionally, if a civilization loses 4 consecutive rounds at any point in the series, the officials begin a fair-play investigation. If the investigation finds evidence of unfair practices, there is a 50% chance that the series is declared void, and will be restarted.

Let G be a random variable denoting the identity of the winning civilization, with possible outcomes $\mathcal{G} = \{\mathcal{A}, \mathcal{M}, \mathcal{S}, \mathcal{N}\}$, for the winner being respectively Andromeda, Milky Way, Sombrero, or none. Compute $H(G)$. **[5 marks]**

- (h) If the series is voided due to fair-play violations, the officials may decide to restart the series. Compute the conditional entropy of the series being restarted, given that a fair-play investigation was initiated, $H(R|F)$. **[3 marks]**

Question 2 [30 marks total]

****All students are expected to attempt this question.**

A standard deck of cards contains 4 *suits* — $\heartsuit, \diamondsuit, \clubsuit, \spadesuit$ (“hearts”, “diamonds”, “clubs”, “spades”) — each with 13 *values* — A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K (The A, J, Q, K are called “Ace”, “Jack”, “Queen”, “King”). Each card has a *colour*: hearts and diamonds are coloured red; clubs and spades are black. Cards with values J, Q, K are called *face cards*.

Each of the 52 cards in a deck is identified by its value V and suit S and denoted VS . For example, $2\heartsuit$, $J\clubsuit$, and $7\spadesuit$ are the “two of hearts”, “Jack of clubs”, and “7 of spades”, respectively. The variable c will be used to denote a card’s colour. Let $f = 1$ if a card is a face card and $f = 0$ otherwise.

1. A card is drawn at random from a thoroughly shuffled deck. Calculate:

- (a) The information $h(c = \text{red}; f = 1)$ in observing a red face card [2 marks]
- (b) The conditional information $h(V = K | f = 1)$ in observing a King given a face card was drawn. [3 marks]
- (c) The entropies $H(S)$ and $H(V, S)$. [3 marks]
- (d) The mutual information $I(V; S)$ between V and S . [3 marks]
- (e) The mutual information $I(V; C)$ between the value and colour of a card using the last result and the data processing inequality [4 marks]

2. Now consider that Twenty cards are removed from a standard deck: All 13 \spadesuit s; the $A\heartsuit$, $2\heartsuit$, $3\heartsuit$, $4\heartsuit$, and $5\heartsuit$; and the $A\diamondsuit$ and $K\diamondsuit$. A card is drawn at random from the reduced deck.

- (a) Calculate the entropies $H(S)$ and $H(V, S)$. [5 marks]
- (b) Calculate $I(V; S)$. Explain why it is different from the $I(V; S)$ when a card is drawn at random from the standard deck of 52 cards (i.e. prior to the removal). [5 marks]
- (c) Calculate $I(V; S | C)$. Give an intuitive explanation as to why knowing the color of a card changes the mutual information between its value and suit. [5 marks]

Question 3 [20 marks total]

****All students are expected to attempt this question.**

- (a) Consider a fair coin flip. What is the mutual information between the top side and the bottom side of the coin **[3 marks]**
- (b) A 6-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)? **[5 marks]**
- (c) Show that the entropy of the probability distribution, $(p_1, \dots, p_i, \dots, p_j, \dots, p_m)$, is less than the entropy of the distribution $(p_1, \dots, \frac{p_i+p_j}{2}, \dots, \frac{p_i+p_j}{2}, \dots, p_m)$. **[6 marks]**
- (d) Let X be a random variable with m different outcomes $\mathcal{X} = \{1, 2, \dots, m\}$ and corresponding nonzero probabilities $\{q_1, q_2, \dots, q_m\}$ arranged in decreasing order. Now define a new random variable Y with $m+1$ different outcomes $\{\frac{q_1}{3}, \frac{q_2}{3}, \dots, \frac{q_{m-1}}{3}, \frac{q_m}{3}, \frac{2}{3}\}$. Find the entropy of $H(Y)$ in terms of $H(X)$. **[6 marks]**

Question 4 [15 marks total]

****Section I: Only COMP 2610 students are expected to attempt this question.**

Let Z be a random variable with possible outcomes 0, 2, and $p(Z = 2) = 1/2$. Let W be a random variable with possible outcomes $W = \{d, e, f\}$. Define

$$\mathbf{r} = (p(W = d|Z = 2), p(W = e|Z = 2), p(W = f|Z = 2))$$

$$\mathbf{s} = (p(W = d|Z = 0), p(W = e|Z = 0), p(W = f|Z = 0)).$$

(a) Suppose that

$$\mathbf{r} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

$$\mathbf{s} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

Compute $I(W; Z)$.

[3 marks]

(b) Using the definition of mutual information, show that for any choice of \mathbf{r}, \mathbf{s} ,

$$I(W; Z) = \frac{D_{\text{KL}}(\mathbf{r}||\mathbf{n} + \mathbf{s}||\mathbf{n})}{2}$$

where $\mathbf{n} = \frac{\mathbf{r} + \mathbf{s}}{2}$.

[5 marks]

(c) Explain the relationship between mutual information, entropy, and conditional entropy. You can use mathematical derivations to explain. **[2 marks]**

(d) Why is mutual information considered a measure of the dependence between two random variables? **[2 marks]**

(e) How does the concept of Kullback-Leibler divergence relate to mutual information? **[3marks]**

****Section II: Only COMP 6261 students are expected to attempt this question.**

Let Y be a random variable with possible outcomes 0, 1, and $p(Y = 1) = 1/2$. Let X be a random variable with possible outcomes $X = \{a, b, c\}$. Define

$$\mathbf{p} = (p(X = a|Y = 1), p(X = b|Y = 1), p(X = c|Y = 1))$$

$$\mathbf{q} = (p(X = a|Y = 0), p(X = b|Y = 0), p(X = c|Y = 0)).$$

- (a) Suppose that

$$\mathbf{p} = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right)$$

$$\mathbf{q} = \left(\frac{1}{5}, \frac{3}{5}, \frac{1}{5}\right)$$

Compute $I(X; Y)$.

[3 marks]

- (b) Using the definition of mutual information, show that for any choice of \mathbf{p}, \mathbf{q} ,

$$I(X; Y) = \frac{D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m} + \mathbf{q} \parallel \mathbf{m})}{2}$$

where $\mathbf{m} = \frac{\mathbf{p} + \mathbf{q}}{2}$.

[3 marks]

- (c) Let Z be a random variable with outcomes $x \in \mathbf{X}$, such that for all $x \in \mathbf{X}$ and $y \in \{0, 1\}$, $p(Z = x|Y = y) = p(X = x|Y = 1 - y)$. Using part (b) compute $I(Z; Y)$. Explain intuitively the relation of your answer to part (b). **[3 marks]**
- (d) Suppose \mathbf{p} and \mathbf{q} are as in part (a). Using part (b), or otherwise, give an example of a random variable Z with possible outcomes \mathbf{X} satisfying $I(X; Y) > I(Z; Y)$. Explain your answer in terms of the data-processing inequality. **[3 marks]**
- (e) Suppose \mathbf{p} and \mathbf{q} are as in part (a). Give an example of a random variable Z with possible outcomes $x \in \mathbf{X}$ satisfying $I(X; Y) < I(Z; Y)$. Explain your answer in terms of the data-processing inequality. **[3marks]**

Question 5 [10 marks total]

****All students are expected to attempt this question.**

Suppose X is a real valued random variable with $\mu = E(X) = 0$

- (a) Show that for any $t > 0$,

$$E(e^{tX}) \leq e^{g(t(b-a))},$$

where $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$, with $\gamma = -a/(b - a)$.

[6marks]

(Hint: write X as a convex combination of a and b , where the convex weighting parameter depends upon X . Exploit the convexity of the function $x \rightarrow e^{tx}$ and the fact that inequalities are preserved upon taking expectations of both sides, since expectations are integrals.)

- (b) By using Taylor's theorem, show that for all $u > 0$,

$$g(u) \leq \frac{u^2}{8}.$$

Furthermore show that

$$E(e^{tX}) \leq e^{\frac{t^2(b-a)^2}{8}}.$$

[4marks]