ASSIGNMENT COVER SHEET

ANU College of Engineering and Computer Science
Australian National University
Canberra ACT 0200 Australia
www.anu.edu.au

+61 2 6125 5254

This coversheet must be attached to the front of your assessment

The assessment is due at 26/09/2022, 5 pm unless otherwise specified in the course outline.

Student ID	u7351505
Student Name	Yifan Luo
Course Code	COMP6261
Course Name	Information Theory
Assignment Item	Assignment 2
Due Date	26th September 2022
Date Submitted	26 th September 2022

I declare that this work:

- ☑ upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Student Academic Integrity;
- ☑ is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;
- ☑ is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- ☑ gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- ☑ in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

Signature	Yifan Luo

COMP6261 Assignment 2. Yifan Luo u7351505

Question 1: Inequalities

1(a) 1. According to the definition of Markov's inequality. we have $P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$ ①

Let X be a random variable of raccoons' height, then $E(X) = 10$.

Reformatting ①, we have $P(X \geq 15) \leq \frac{10}{15} = \frac{2}{3} \approx 0.67$

2. According to the definition of Chebyshev's inequality. we have $P(|X - E(X)| \geq \lambda) \leq \frac{Var(X)}{\lambda^2}$. ①

Let X be a random variable of raccoons' height, then $Var(X) = (Std(X))^2 = 2^2 = 4$, $E(X) = 10$

Then by reformatting ①, we have $P(|X - 10| \geq \lambda) \leq \frac{4}{\lambda^2}$ ②

$LHS = 1 - P(10 - \lambda < X < 10 + \lambda) \leq \frac{4}{\lambda^2}$.

Let $\lambda = 5$, we have $1 - P(5 < X < 15) \leq \frac{4}{25}$

Then, $P(5 < X < 15) \geq 1 - \frac{4}{25} = \frac{21}{25} = 0.84$

1(b) 1. Let X be a random variable of head counts in N times flipping.

Then $X \sim B(N, p)$, where $p < \frac{1}{6}$. We have $E(X) = Np$

By using Markov's inequality, we have $P(X \geq \frac{N}{3}) \leq \frac{Np}{\frac{N}{3}} = 3p < \frac{3}{6} = \frac{1}{2}$

2. Based on the above definition, we have $Var(X) = Np(1-p)$

According to the definition of Chebyshev's inequality. we have $P(|X - E(X)| \geq \lambda) \leq \frac{Var(X)}{\lambda^2}$.

$LHS = 1 - P(E(X) - \lambda < X < E(X) + \lambda) \leq \frac{Var(X)}{\lambda^2}$ ①

In ①, Let $E(X) - \lambda = 0$ ②, $E(X) + \lambda = \frac{N}{3}$ ③,

Solve ② and ③, we have $\lambda = E(X) = \frac{N}{3} - E(X)$, where $E(X) = Np$

Reformatting ①, $P(0 < X < \frac{N}{3}) \geq 1 - \frac{Var(X)}{\lambda^2} = 1 - \frac{Np(1-p)}{(\frac{N}{6})^2} = 1 - \frac{36Np(1-p)}{N^2}$

$RHS = 1 - \frac{36Np(1-p)}{N^2} = 1 - \frac{36p(1-p)}{N} > 1 - \frac{6(1-p)}{N} > 1 - \frac{6}{N}$, where $p < \frac{1}{6}$.

Then $P(0 < X < \frac{N}{3}) > 1 - \frac{6}{N}$

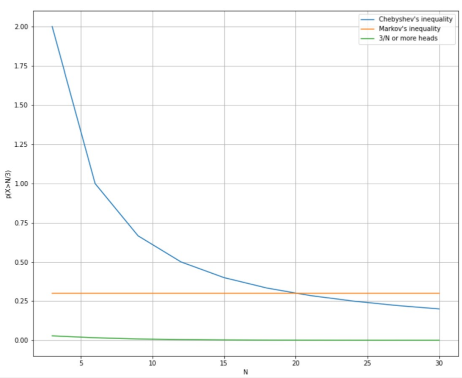
Thus $P(X > \frac{N}{3}) = 1 - P(0 < X < \frac{N}{3}) < \frac{6}{N}$

3. Let $p = \frac{1}{10}$, Python code and plot:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom

N = np.arange(3, 30 + 3, 3)
p = 1 / 10
cdf = binom.cdf(N / 3, N, p)

plt.figure(figsize=(12, 10))
plt.plot(N, 6 / N, label="Chebyshev's inequality")
plt.plot(N, np.repeat(3 * p, len(N)), label="Markov's inequality")
plt.plot(N, 1 - cdf, label="3/N or more heads")
plt.xlabel("N")
plt.ylabel("p(X>N/3)")
plt.legend()
plt.grid()
plt.show()
```



Question 2. Markov Chain

2(a) 1. Yes. Suppose X, Y and Z are independent with each other, then have $p(y)=p(y|x)$, $p(z)=p(z|y)$

(X, Y, Z) forms a Markov Chain, because $p(x,y,z)=p(x)\cdot p(y)\cdot p(z)=p(x)\cdot p(y|x)\cdot p(z|y)$

$\because X\perp Y,\ X\perp Z \quad \therefore p(x|y)=p(x),\ p(x|z)=p(x)$

$\therefore H(X|Y)=\sum_y p(y)H(X|Y=y)=\sum_y p(y)\sum_x p(x|y)\log\frac{1}{p(x|y)}=\sum_y p(y)\sum_x p(x)\log\frac{1}{p(x)}=\sum_y p(y)H(X)=H(X)$

$H(X|Z)=\sum_z p(z)H(X|Z=z)=\sum_z p(z)\sum_x p(x|z)\log\frac{1}{p(x|z)}=\sum_z p(z)\sum_x p(x)\log\frac{1}{p(x)}=\sum_z p(z)H(x)=H(X)$

$\therefore I(X;Y)=H(X)-H(X|Y)=H(X)-H(X)=0$

$I(X;Z)=H(X)-H(X|Z)=H(X)-H(X)=0$

$\therefore I(X;Y)=I(X;Z)=0$

2. Yes. Suppose Y and Z are independent, but X and Z are not independent. We have $p(z)\neq p(z|x)$

(X, Y, Z) can not form a Markov Chain, because $p(x,y,z)=p(x)\cdot p(y|x)\cdot p(z|x,y)$,

and $p(z|y)=p(z)\neq p(z|x)=p(z|x,y)$, then $p(x,y,z)\neq p(x)\cdot p(y|x)\cdot p(z|y)$.

$\because Y\perp Z \quad \therefore I(Y;Z)=H(Y)-H(Y|Z)=H(Y)-H(Y)=0$

According to the non-negativity of Mutual Information, for any X and Y

we have $I(X;Y)\geqslant 0=I(Y;Z)$

3. Before the prove of the two inequalities, we first prove the chain rule of Mutual Information.

According to the definition, $I(X;Y,Z)=H(X)-H(X|Y,Z)$ ①

$RHS=H(X)-H(X|Y)+H(X|Y)-H(X|Y,Z)=I(X;Y)+I(X;Z|Y)$ ②

Also, $RHS=H(X)-H(X|Z)+H(X|Z)-H(X|Y,Z)=I(X;Z)+I(X;Y|Z)$ ③

Based on ①, ② and ③, we have $I(X;Y,Z)=I(X;Y)+I(X;Z|Y)$ ④

$=I(X;Z)+I(X;Y|Z)$ ⑤

Then ④ = ⑤, we have $I(X;Y)+I(X;Z|Y)=I(X;Z)+I(X;Y|Z)$ ⑥

$I(X;Z|Y)=D_{KL}(p(x,z|y)\,\|\,p(x|y)p(z|y))=\sum_x\sum_y\sum_z p(x,z|y)\log\frac{p(x,z|y)}{p(x|y)p(z|y)}$

$\because X\to Y\to Z \quad \therefore \frac{p(x,z|y)}{p(x|y)p(z|y)}=\frac{p(x,y,z)}{p(y)p(x|y)p(z|y)}=\frac{p(x,y,z)}{p(x)p(y|x)p(z|y)}=\frac{p(x,y,z)}{p(x,y,z)}=1$

$\therefore I(X;Z|Y)=\sum_x\sum_y\sum_z p(x,z|y)\log 1=0$

Reformating ⑥, we have $I(X;Y)=I(X;Z)+I(X;Y|Z)-I(X;Z|Y)=I(X;Z)+I(X;Y|Z)$ ⑦

☆ Then, $I(X;Z)\leq I(X;Y)$, the first inequality is proved.

From ⑦, we have $I(X;Y|Z)=I(X;Y)-I(X;Z)$

☆ Then, $I(X;Y|Z)\leq I(X;Y)$, the second inequality is proved.

2(b) 1. According to the definition, $I(X;Y,Z,T) = H(X) - H(X|Y,Z,T)$ ①

Since we already prove the chain rule of Mutual Information in 2(a),

by reformatting ①, we have $I(X;Y,Z,T) = I(X;Y,Z) + I(X;T|Y,Z)$ ②

$= I(X;T) + I(X;Y,Z|T)$ ③

② = ③, then $I(X;Y,Z) + I(X;T|Y,Z) = I(X;T) + I(X;Y,Z|T)$ ④

Same as 2(a), we first prove $I(X;T|Y,Z) = 0$ by following below

$$I(X;T|Y,Z) = D_{KL}(p(X,T|Y,Z)\|p(X|Y,Z)p(T|Y,Z))$$

$$= \sum_y\sum_z\sum_x\sum_T p(x,T|Y,Z)\log\frac{p(X,T|Y,Z)}{p(X|Y,Z)p(T|Y,Z)}$$

$\because X \to (Y,Z) \to T$ $\quad\therefore P(X,Y,Z,T) = p(x)p(y,z|x)p(T|Y,Z)$

$$\therefore \frac{p(X,T|Y,Z)}{p(X|Y,Z)p(T|Y,Z)} = \frac{p(X,T,Y,Z)}{p(Y,Z)p(X|Y,Z)p(T|Y,Z)} = \frac{p(X,T,Y,Z)}{p(x)p(Y,Z|x)p(T|Y,Z)} = \frac{p(X,T,Y,Z)}{p(X,T,Y,Z)} = 1$$

$\therefore I(X;T|Y,Z) = 0$

Then, reformatting ④, we have $I(X;Y,Z) = I(X;T) + I(X;Y,Z|T)$ ⑤

$\therefore I(X;Y,Z) \geq I(X,T)$

2. From the above ⑤, we have $I(X;Y,Z) = I(X;T) + I(X;Y,Z|T)$

If $I(X;Y,Z) = I(X,T)$, then $I(X;Y,Z|T) = 0$.

$I(X;Y,Z|T) = I(X;Y,Z,T) - I(X;T) = 0 \Rightarrow I(X;Y,Z,T) = I(X;T)$

$LHS = H(X) + H(Y,Z,T) - H(X,Y,Z,T)$

$= H(X) + H(T) + H(Y,Z|T) - (H(X,T) + H(Y,Z|X,T))$

$= H(X) + H(T) - H(X,T) + H(Y,Z|T) - H(Y,Z|X,T)$

$RHS = H(X) + H(T) - H(X,T)$

If $H(Y,Z|T) = H(Y,Z|X,T) = 0$, $LHS = RHS$, i.e. $I(X;Y,Z) = I(X;T)$

Thus the condition is given T, (Y,Z) and X are independent.

2(c) 1. Suppose $p(X=a) = 1$, otherwise $p(x) = 0$

Then, $E(x) = \sum_x xp(x) = a \quad (\forall a > 0)$

We have $P(x \geq a) = \frac{E(x)}{a} = 1$

2. Suppose Y is uniformlly distributed in range $[-1,1]$. Then $E(Y) = 0$.

We have $\forall a \geq 0$, $P(Y \geq a) \geq \frac{E(Y)}{a} = 0$,

because any probability is non-negative.

3. Yes. Suppose $P(Z=1)=P(Z=-1)=\frac{1}{2}$, otherwise $p(z)=0$.

We have $E(Z)=1\cdot\frac{1}{2}+(-1)\cdot\frac{1}{2}=0$, $E(Z^2)=1^2\cdot\frac{1}{2}+(-1)^2\cdot\frac{1}{2}=1$

Then, $E(|Z|)=|-1|\cdot\frac{1}{2}+1\cdot\frac{1}{2}=1$, $Var(Z)=E(Z^2)-(E(Z))^2=1-0=1$

Suppose $a=0.1$, then $\frac{E(|Z|)}{a}=10\,E(|Z|)=10$, $\frac{Var(Z)}{a^2}=100\,Var(Z)=100$

Then we have $\frac{E(|Z|)}{a}<\frac{Var(Z)}{a^2}$

Question 3. AEP

a) $H(X)=E(h(x))=p_h\cdot\log_2\frac{1}{p_h}+p_t\cdot\log_2\frac{1}{p_t}$

$=-0.8\cdot\log_2 0.8-0.2\log_2 0.2$

$\simeq 0.72$

b) $|A_{X^N}|=2^N$

c) $H_0(X^4)=\log_2|A_{X^4}|=\log_2 2^4=4$ bits

d) We know that entropy is additive for independent random variables. in this case, $x\in X^N$ is drawn i.i.d. according to $P_X=\{p_h,p_t\}$, then we have

$H(X^N)=N\,H(X)=N(-p_h\log_2 p_h-p_t\log_2 p_t)=N(-0.8\log_2 0.8-0.2\log_2 0.2)\simeq 0.72N$

e)

```
import numpy as np
import matplotlib.pyplot as plt
from math import log2
from scipy.special import comb
```

executed in 711ms, finished 15:03:30 2022-09-26

```
p_h, p_t = 0.8, 0.2
N = [10, 50]
deltas = np.linspace(0, 1, num=50, endpoint=True)
```

executed in 14ms, finished 15:03:30 2022-09-26

```
def essential_bit_content(p, N, deltas):
    if 1 - p > p:
        p = 1 - p

    probs = []
    for k in range(N + 1):
        p_N_k = (p ** k) * ((1 - p) ** (N - k))
        probs.extend([p_N_k for _ in range(int(comb(N, k)))])  # ascending order

    avg_essential_bit_contents = []
    for delta in deltas:
        while sum(probs[1:]) > 1 - delta:
            probs.pop(0)
        avg_essential_bit_contents.append(log2(len(probs)) / N)

    return deltas, avg_essential_bit_contents


def plot_essential_bit_content(D, H, N):
    plt.figure(figsize=(10, 6))
    plt.scatter(D, H, s=0.1)
    for i in range(len(D) - 1):
        xs = np.linspace(D[i], D[i + 1], endpoint=True)
        plt.plot(xs[:-1], np.repeat(H[i], len(xs) - 1), 'r')
        ys = np.linspace(H[i], H[i + 1], endpoint=True)
        plt.plot(np.repeat(D[i + 1], len(ys) - 1), ys[:-1], 'r')
```

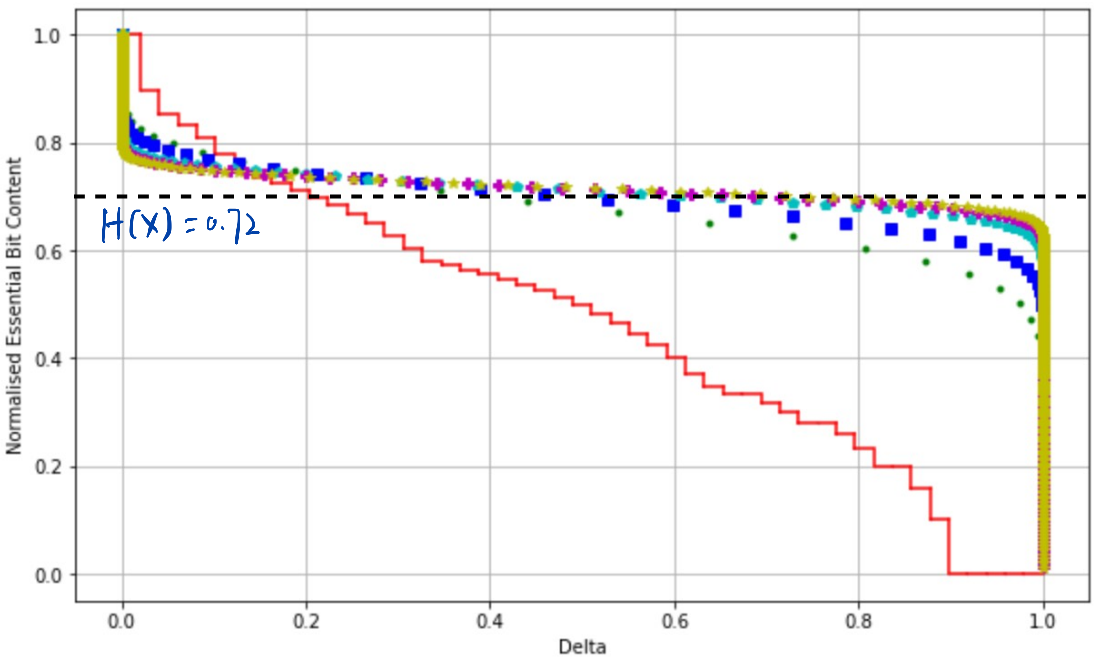
executed in 14ms, finished 15:03:30 2022-09-26

```python
N = [100, 200, 500, 800, 1010]
colors = ['g', 'b', 'c', 'm', 'y']
markers = ['.', 's', 'p', 'P', '*']


def plot_essential_bit_content2(p, N, c, m):
    n_total = 0
    for i in range(N):
        n_total += int(comb(N, i))
    if 1 - p > p:
        p = 1 - p
    delta = 0
    for k in range(N + 1):
        p_N_k = (p ** k) * ((1 - p) ** (N - k))
        n_p_N_k = int(comb(N, k))
        delta += p_N_k * n_p_N_k
        n_total -= n_p_N_k
        if n_total == 0:
            break
        avg_essential_bit_content = log2(n_total) / N
        plt.plot(delta, avg_essential_bit_content, c=c, marker=m)


D, H = essential_bit_content(p_h, 10, deltas)
plot_essential_bit_content(D, H, 10)
for n, c, m in zip(N, colors, markers):
    plot_essential_bit_content2(p_h, n, c, m)
plt.grid()
plt.xlabel("Delta")
plt.ylabel("Normalised Essential Bit Content")
```

executed in 1.98s, finished 15:03:32 2022-09-26



The picture above illustrates $\frac{1}{N}H_\delta(X^N)$ v.s. δ when $N = \{10, 100, 200, 500, 800, 1010\}$, colored in 6 base colors of matpotlib with different shapes.

① The curve is more flattern for large N.

② Given a tiny probability of δ, $\frac{1}{N}H_\delta(X^N)$ can be made as close to H as required. For a large probability of δ, we can not compress more than H bits per outcome for large sequences.

③ As we increase δ, we will quickly encounter typical sets that make small, roughly equal sized changes to $|S_\delta|$

④ It derives the Asymptotic Equipartition Property and Source coding theorem.

Question 5. AEP

1. (a) $\log_2 4 = 2$ bits e.g. $\{00, 01, 10, 11\}$ for Alina, Beyonce, Cecilia and Derek

(b) $\log_2 7 < \log_2 8 = 3$ bits e.g. $\{000, 001, 010, 011, 100, 101, 110\}$ for Track 1 to Track 7

(c) $\log_2(7+12+15+14) = \log_2 48 < \log_2 64 = 6$ bits

2. Suppose $\mathcal{A}_X$ is a set of all the tracks, then $|\mathcal{A}_X| = 7+12+15+14 = 48$
Raw bit content of X: $H_0(X) = \log_2|\mathcal{A}_X| = \log_2 48 \approx 5.58$

3. (a) $\mathcal{A}_A = \{$Alina, Beyonce, Cecilia, Derek$\}$
$\mathcal{P}_A = \{P(A=\text{Alina}) = \frac{7}{48}, P(A=\text{Beyonce}) = \frac{1}{4}, P(A=\text{Cecilia}) = \frac{5}{16}, P(A=\text{Derek}) = \frac{7}{24}\}$
Thus, the ensemble of A is $(a, \mathcal{A}_A, \mathcal{P}_A)$

(b) $H_0(A^4) = \log_2|\mathcal{A}_{A^4}| = \log_2 4^4 = 8$ bits

(c) When $\delta = 0$, then $|S_\delta| = 4^4 = 256$.
Remove one element with the lowest probability, i.e. (Alina, Alina, Alina, Alina)
The probability is $(\frac{7}{48})^4$, then $|S_\delta \setminus (\text{Alina, Alina, Alina, Alina})| = 255 < 256$
Thus, the smallest value of δ is $(\frac{7}{48})^4 \approx 0.00045$

(d) The top 2 elements with largest probabilities are (Cecilia, Cecilia, Cecilia, Cecilia), ①
and (Cecilia, Cecilia, Cecilia, Derek) ② (the order doesn't matter)
The propability of ① is $(\frac{5}{16})^4$, the probability of ② is $(\frac{5}{16})^3 \cdot \frac{7}{24}$
Let $\delta = 1 - (\frac{5}{16})^4 - (\frac{5}{16})^3 \cdot \frac{7}{24}$, then $H_\delta(|S_\delta|) = \log_2 2 = 1 > 0$
Thus, the largest value of δ is $1 - (\frac{5}{16})^4 - (\frac{5}{16})^3 \frac{7}{24} \approx 0.98$

4. (a) $H(A) = -\sum_{a \in \mathcal{A}_A} p(a) \log(p(a))$, where $\mathcal{A}_A = \{$Alina, Beyonce, Cecilia, Derek$\}$

$$= -\frac{7}{48}\log_2\frac{7}{48} - \frac{1}{4}\log_2\frac{1}{4} - \frac{5}{16}\log_2\frac{5}{16} - \frac{7}{24}\log_2\frac{7}{24}$$
$$\approx 1.948$$

Then $H(A^4) = 4H(A) \approx 7.79$, and $H(A) \approx 1.95$ in case you need.

(b) According to, $|T_{N\beta}| \le 2^{N(H(A)+\beta)}$, we have $N=100$, $\beta = 0.1$

From (a), we have $H(A) = 1.95$, then $|T_{N\beta}| \le 2^{100(1.95+0.1)}$

(c) In lecture 12, we mentiond that " Given a tiny probability of erro δ, the average bits per outcome can be made as close to H as required."

In this case, we have $\delta = 1 - 95\% = 0.05$, and $H = H(A) \approx 1.948$.

Then, when $n \to \infty$, $\frac{1}{N} H_\delta(A^N) \to 1.948$, which is larger than 1.5,

Thus, it is not possible.