EE 231A Shengze Ye 205418959 HW1

1. Coin Flips

$H(p) = -p\log p - q\log q \qquad p = 1-q$

Suppose a coin with probability of head p is flipped until the first head occurs. X denote the number of flips required

(a) Find the entropy $H(X)$

$P(X=1) = p \quad p(X=2) = pq \quad p(X=3) = pq^2$

$\cdots \quad p(X=n) = pq^{n-1}$

$$H(X) = -\sum_{n=1}^{\infty} pq^{n-1}\log(pq^{n-1})$$

$$= -p\log p - pq\log pq - pq^2\log pq^2 - \cdots - pq^{n-1}\log pq^{n-1} - \cdots$$

$$= -p\log p - pq[\log p + \log q] - pq^2[\log p + 2\log q] -$$

$$pq^3[\log p + 3\log q] - \cdots - pq^{n-1}[\log p + (n-1)\log q] - \cdots$$

$$= -p\log p \sum_{n=0}^{\infty} q^n - p\log q \sum_{n=1}^{\infty} nq^n$$

consider $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad \sum_{n=1}^{\infty} nq^n = r/(1-r)^2$

$$\therefore H(X) = -p\log p \frac{1}{1-q} - p\log q \frac{q}{(1-q)^2}$$

$\because\ p+q=1 \qquad \therefore\ 1-q=p$

$$H(X) = -p\log p \cdot \frac{1}{p} - p\log q \cdot \frac{q}{p^2}$$

$$= \frac{1}{p}\left[-p\log p - q\log q\right]$$

$$= \frac{1}{p} H(p) \text{ bits}$$

$$\therefore H(X) = \frac{1}{p} H(p) \text{ bits}$$

(b) What is the value for a fair coin?

For a fair coin, $p=q=\frac{1}{2}$

$$H(X) = 2 \times H\left(\frac{1}{2}, \frac{1}{2}\right) = 2\times 1 = 2$$

$H(X) = 2$

(c) A random variable X is drawn according to the distribution with $p=\frac{1}{2}$

$\rightarrow H(X) = 2$

$P(X=1)=\frac{1}{2} \quad P(X=2)=\frac{1}{4} \quad \cdots \quad P(X=n)=\frac{1}{2^n}$

One possible "efficient" sequence yes-no question is:

① Is $X=1$? if not ↙

② Is $X=2$? if not ↙

③ Is $X=3$? ↙

⋮

So the pattern is if $X = r$, we need to ask r questions

And the expected number of questions should be

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \cdots + \frac{1}{2^n} \cdot n$$

$$= \sum_{n=1}^{\infty} n \cdot \frac{1}{2^n} = 2$$

$\therefore H(X)$ is equal to the expected number of questions required to determine X

2. Minimum Entropy

$H(p_1, p_2, \ldots p_n) = H(\vec{p})$

According to the property of entropy, $H(p)$ is a concave function of the distribution.

For $H(\vec{p})$, it achieves its minimum value as $p_i = 1$ for some i and $p_j = 0$ for $j \neq i$

e.g. $p = (1, 0, \ldots 0)$ $p = (0, 1, 0, \ldots 0)$

$\ldots$ $p = (0, 0, \ldots, 0, 1)$

Since $p = 0$ or 1, the variable is not random, and there is no uncertainty. So $H(\vec{p}) = 0$

These points correspond to corners on the simplex

3. Entropy of function of a random variable

(a) $H(X, g(X)) = H(X) + H(g(X)|X)$

$$= -\sum_x p(x) \log p(x) - \sum_x \sum_{g(x)} p(x, g(x)) \log p(g(x)|x)$$

$$= -\sum_x p(x) \sum_{g(x)} p(g(x)|x) \log p(x) - \sum_x \sum_{g(x)} p(x, g(x)) \log p(g(x)|x)$$

$$= -\sum_x \sum_{g(x)} p(x, g(x)) \left[\log p(x) + \log p(g(x)|x) \right]$$

$$= -\sum_x \sum_{g(x)} p(x, g(x)) \log \left(p(x) \cdot p(g(x)|x) \right)$$

$$= -\sum_x \sum_{g(x)} p(x, g(x)) \log p(x, g(x)) = H(x, g(x))$$

(b)

$H(X) + H(g(x)|X)$

since $g(x)$ is a function of x, when x is fixed

$g(x)$ is fixed ... there is no uncertainty in $p(g(x)|x)$

... $H(X) + H(g(x)|X) = H(X) + 0 = H(X)$

(c) $H(X, g(x)) = H(g(x)) + H(X|g(x))$

Can be proved by chain rule, similar to (a)

(d) $H(g(x)) + H(x|g(x))$

We know that $H(x|g(x)) \geq 0$

if $x \to g(x)$ is one-to-one mapping, we know x if we already know $g(x)$. if it is not one-to-one, the is uncertainty in $p(x|g(x))$

$\therefore H(x|g(x)) \geq 0$

$H(g(x)) + H(x|g(x)) \geq H(g(x))$

$\therefore H(x, g(x)) = H(x) \geq H(g(x))$

4. computing Relative Entropy for 2D p and q

$p(x,y)$

$x \backslash Y$	0	1
0	$\frac{1}{6}$	$\frac{7}{12}$
1	$\frac{1}{6}$	$\frac{1}{12}$

$q(x,y)$

$x \backslash Y$	0	1
0	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{12}$	$\frac{1}{6}$

Find $D(p||q)$

$$D(p||q) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

$$= \frac{1}{6} \log \frac{\frac{1}{6}}{\frac{1}{4}} + \frac{7}{12} \log \frac{\frac{7}{12}}{\frac{1}{2}} + \frac{1}{6} \log \frac{\frac{1}{6}}{\frac{1}{12}}$$

$$+ \frac{1}{12} \log \frac{\frac{1}{12}}{\frac{1}{6}}$$

$= \frac{1}{6}\log\frac{2}{3} + \frac{7}{12}\log\frac{7}{6} + \frac{1}{6}\log 2 + \frac{1}{12}\log\frac{1}{2}$

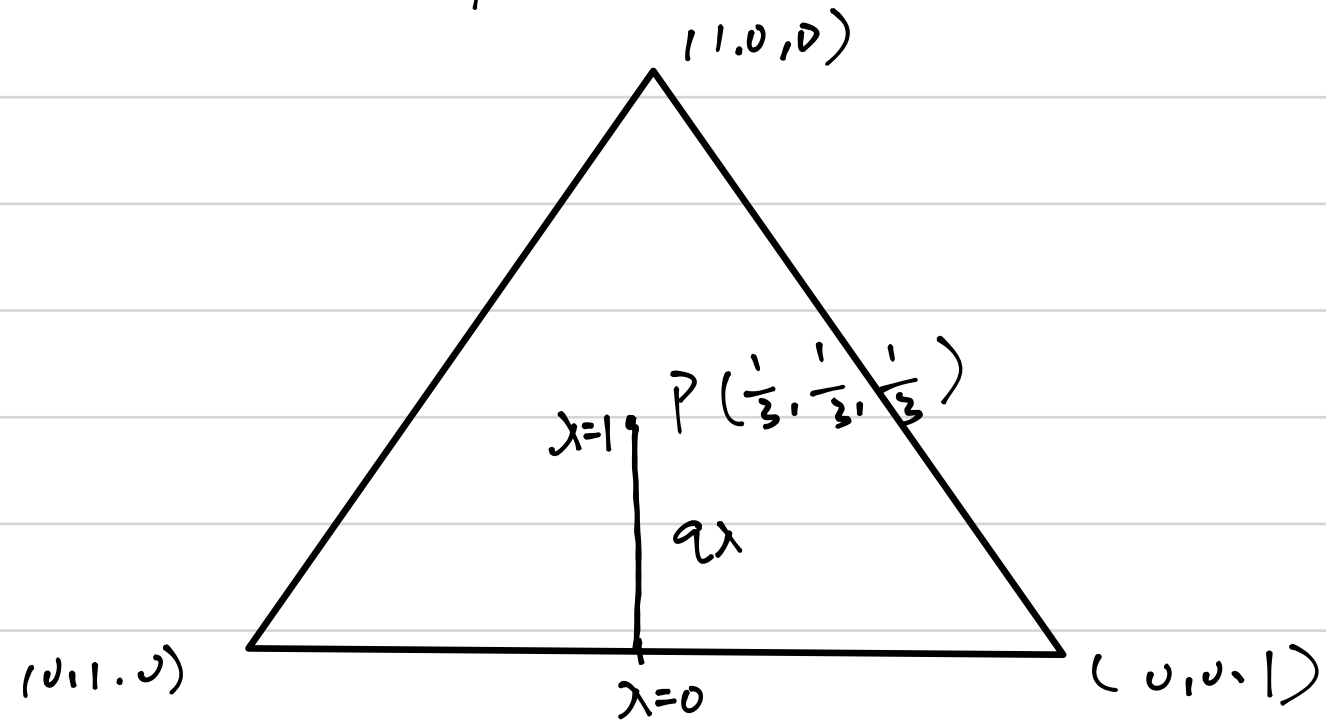
$= 0.1156$ bit $\quad \therefore D(p\|q) = 0.1156$ bit

5. Computing Relative Entropy for p and q on a line in 3D simplex.

PMF of p is point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

PMF of q_λ is point $(\frac{\lambda}{3}, \frac{3-\lambda}{6}, \frac{3-\lambda}{6})$

(a) draw a triange representing the simplex, show the point p and the line segment the shows the trajectory of q as λ varies from 0 and 1

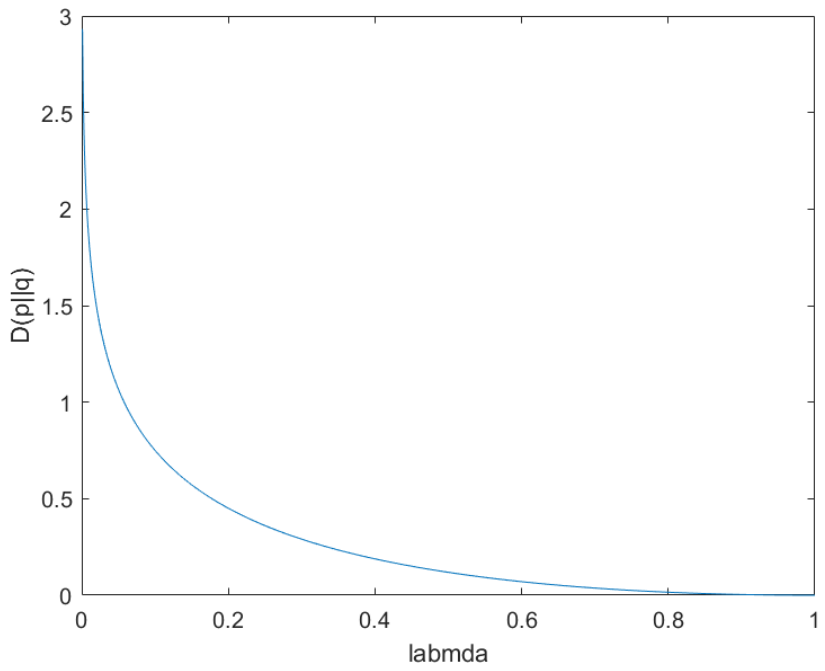


when $\lambda = 0$ $q = (0, \frac{1}{2}, \frac{1}{2})$ when $\lambda = 1$ $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

as λ varies from 0 to 1, q varies from the middle of the side to P

(b) Find $D(p\|q_\lambda)$ as a function of λ as λ varies between 0 and 1 use MATLAB

$D(p\|q_\lambda)$



when $\lambda = 1$ p and q represent the same distribution.

$\therefore D(p\|q) = 0$

when $\lambda = 0$ $q = (0, \frac{1}{2}, \frac{1}{2})$

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \frac{1}{3}\log\frac{\frac{1}{3}}{0} + \frac{1}{3}\log\frac{\frac{1}{3}}{\frac{1}{2}} + \frac{1}{3}\log\frac{\frac{1}{3}}{\frac{1}{2}}$$

As defined in the book, we consider $p\log\frac{p}{0} = \infty$

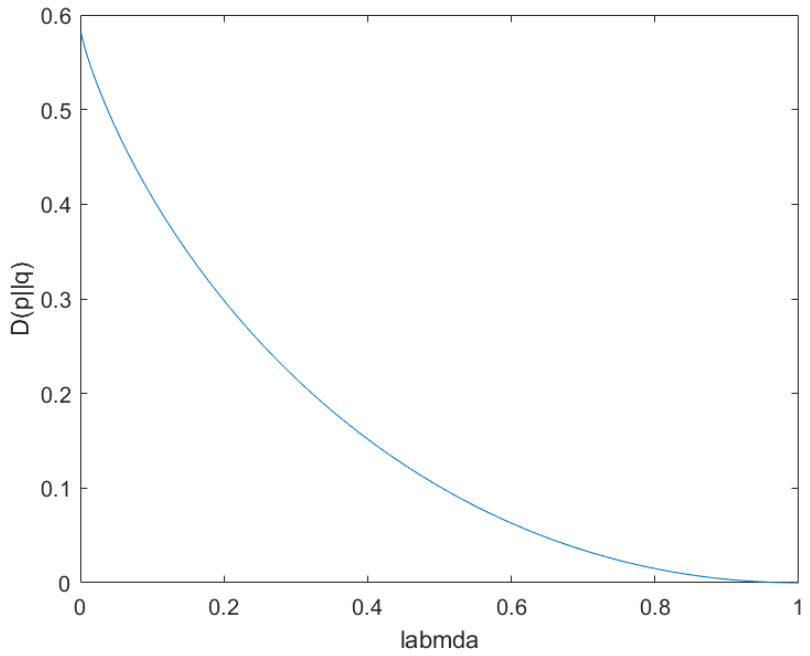
$$\therefore D(p\|q) = +\infty + \frac{1}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{2}{3}$$

$$= +\infty$$

$\therefore$ when $\lambda = 0$ $D(p\|q)$ reaches positive definite

(c) Find $D(q_\lambda \| p)$ as a function of λ as λ varies between 0 and 1 use MATLAB

$D(q_\lambda \| p)$



Again when $\lambda = 1$, p, q represent the same distribution

$\therefore D(q \| p) = 0$

when $\lambda = 0$

$$D(q \| p) = 0 \log \frac{0}{\frac{1}{3}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{3}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{3}}$$

as defined in the book, we consider $0 \log \frac{0}{q} = 0$

$$\therefore D(q \| p) = \left(\frac{1}{2} \log \frac{3}{2}\right) \times 2 = \log \frac{3}{2}$$

(As $\log_2 0$ is not defined, we cannot get the point when $\lambda = 0$ in MATLAB)

(d) Discuss the differences between $D(p||q_\lambda)$ and $D(q_\lambda||p)$

$D(p||q)$ is the penalty when p is the true distribution, q is the wrong distribution

On the contrary, $D(q||p)$ is the penalty when q is the true distribution, p is the wrong distribution

We note that the "distance" between p and q is not symmetric as $D(p||q) \neq D(q||p)$ when $\lambda = 0$

$p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $q = (0, \frac{1}{2}, \frac{1}{2})$

For $D(p||q)$

The $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is true distribution, but we wrongly set one variable to 0 possibility, this leads to the infinite penalty

For $D(q||p)$

The $q = (0, \frac{1}{2}, \frac{1}{2})$ is true distribution, but we wrongly use $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. the penalty for this situation is smaller than the previous situation.

6. Mutual Information ?

Can the relative entropy computed in problem 4 be express as a mutual information ?

Yes.

The definition of $I(X;Y)$ is

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

For problem 4 $D(p\|q) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)}$

$p(x,y)$:

$x \backslash Y$	0	1
0	$\frac{1}{6}$	$\frac{7}{12}$
1	$\frac{1}{6}$	$\frac{1}{12}$

$\therefore p(X=0) = \frac{3}{4}$ $p(X=1) = \frac{1}{4}$

$p(Y=0) = \frac{1}{3}$ $p(Y=1) = \frac{2}{3}$

$q(x,y)$:

$x \backslash Y$	0	1
0	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{12}$	$\frac{1}{6}$

$q(0,0) = p(X=0) \cdot p(Y=0) = \frac{1}{4}$

$q(0,1) = p(X=0)\,p(Y=1) = \frac{1}{2}$

$q(1,0) = p(X=1)\,p(Y=0) = \frac{1}{12}$

$q(1,1) = p(X=1)\,p(Y=1) = \frac{1}{6}$

$\therefore q(x,y) = p(x) \cdot p(y)$

$\therefore D(p\|q) = \sum_{x,y} p(x,y) \log p(x,y)/q(x,y)$

$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$

$\therefore$ It could be expressed as a mutual information

7. Example of joint entropy

$p(x,y)$:

$x \backslash y$	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

(a) $P(X=0) = \frac{2}{3}$ $\quad P(X=1) = \frac{1}{3}$

$H(X) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$ bit

$P(Y=0) = \frac{1}{3}$ $\quad P(Y=1) = \frac{2}{3}$

$\therefore H(Y) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$ bit

$H(X) = H(Y) = 0.918$ bit

(b) $H(X|Y) = \sum_y p(y)\, H(X|Y=y)$

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

$= \frac{1}{3} H(1,0) + \frac{2}{3} H(\frac{1}{2},\frac{1}{2}) = \frac{2}{3} = 0.667$ bit

$H(Y|X) = \sum_x p(x)\, H(Y|X=x)$

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

$= \frac{2}{3} H(\frac{1}{2},\frac{1}{2}) + \frac{1}{3} H(0,1) = \frac{2}{3} = 0.667$ bit

(c) $H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y)$

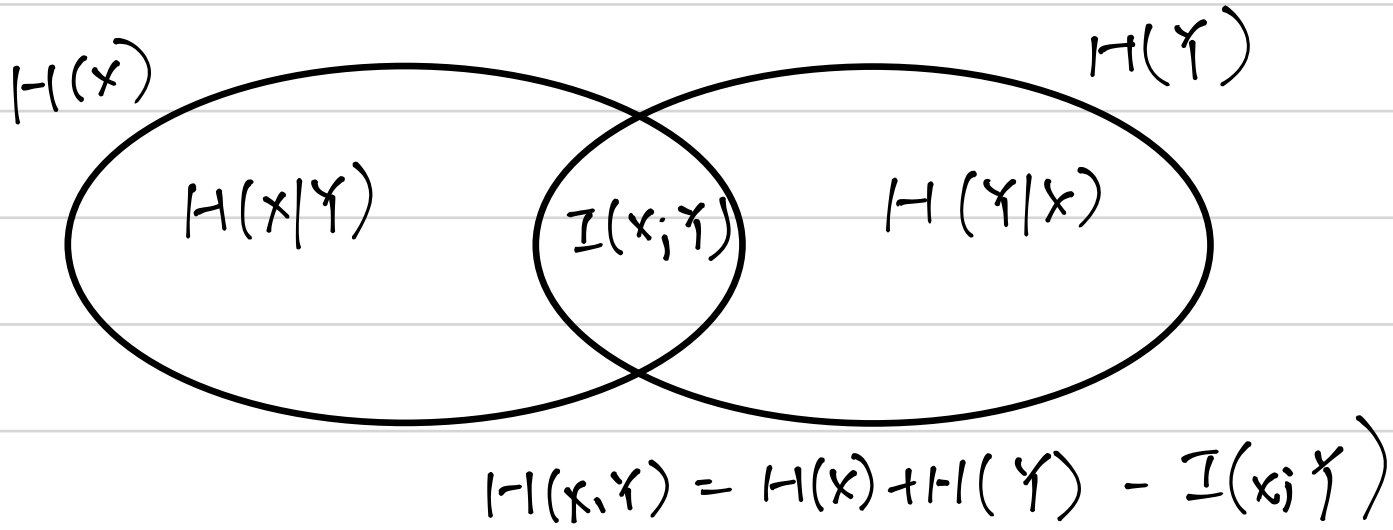
$= -3 \times \frac{1}{3}\log\frac{1}{3} = \log 3 = 1.58$ bits

(d) $H(Y) - H(Y|X)$

$= 0.918\ bit - 0.667\ bit = 0.251\ bit$

(e) $I(X;Y) = H(Y) - H(Y|X) = 0.251\ bit$

(f) Draw a Venn diagram for the quantities above



$H(X,Y) = H(X) + H(Y) - I(X;Y)$

8. Mutual Information and the Weather

Let random variable W denote the weather in El Nino University

according to the question

$P(W = rain) = 0.9 \quad p(W = sun) = 0.1$

(a) Stormy is a very lazy forecaster, it achieves his 90% correct forecast by simply always predict rain

∴ Let random variable P_S denote the predicted weather for stormy forecaster

∴ $p(P_S = sun) = 0 \quad p(P_S = rain) = 1$

Therefore, we can derive the join distribution

$p(Ps, W)$:

Ps \ W	sun	rain
sun	0	0
rain	0.1	0.9

$$I(Ps; W) = H(W) - H(W|Ps)$$

$$H(W) = H(0.1, 0.9)$$

$$H(W|Ps) = \sum_{Ps} p(ps)\, H(W|Ps = ps)$$

$$= 0 + 1 \cdot H(0.1, 0.9) = H(0.1, 0.9)$$

$$\therefore I(Ps; W) = H(W) - H(W|Ps) = 0 \text{ bit}$$

(b) Wendy is a hard-working forecaster.

Let random variable Pw denote the predicted weather for wendy forecaster

According to the (b) $p(Pw = sun) = 0.25$

$p(Pw = rain) = 0.75$

Whenever she predicts rain she is correct

$\therefore p(W = rain \mid Pw = rain) = 1$

$$p(W = rain, pw = rain) = p(W = rain \mid Pw = rain) \cdot p(Pw = rain)$$

$$= 1 \cdot 0.75 = 0.75$$

Whenever she predicts sun she is correct 40% of the time

$\therefore P(W=\text{sun} \mid P_W=\text{sun}) = 0.4$

$P(W=\text{sun}, P_W=\text{sun}) = P(W=\text{sun} \mid P_W=\text{sun}) \cdot P(P_W=\text{sun})$

$= 0.4 \cdot 0.25 = 0.1$

$\therefore$ we can draw the join distribution here

$P_W \backslash W$	sun	rain
sun	0.1	0.15
rain	0	0.75

$I(p_W; W) = H(W) - H(W \mid P_W)$

$H(W) = -0.1 \log 0.1 - 0.9 \log 0.9 = 0.469$ bit

$H(W \mid P_W) = \sum_{p_W} P(p_W)\, H(W \mid P_W = p_W)$

$= 0.25 \cdot \underset{①}{H(0.6, 0.4)} + 0.75\, \underset{②}{H(0.1)}$

① $H(0.6, 0.4) = -0.6 \log 0.6 - 0.4 \log 0.4$

$= 0.971$

② $H(0.1) = 0$

$\therefore H(W \mid P_W) = 0.2427$ bit

$\therefore I(p_W; W) = 0.469 - 0.2427 = 0.2263$ bit

(c) $I(P_S; W) = 0$ bit, $I(p_W; W) = 0.2263$ bit

$\therefore$ Wendy forecaster provides the most information about the weather

(d) Suppose you want to plant your tulip bulbs on a day when you know it was going to rain. which forecaster would be helpful in achieving this?

I will select Wendy forecast, as the Storm forecaster always predict rain. it provides no information to me, it is helpless.

For forecaster Wendy, when she predicts rain. it will rain, this will help me in this situation

9. Concavity of entropy

(a) Show that $\log x$ is concave in x for positive x

$\log x = \ln x / \ln 2$

$\ln 2$ is a constant, if the base for log is b,

$\log_b x = \ln x / \ln b$. it should be a constant

$$\frac{d}{dx}\log x = \frac{d}{dx}\frac{1}{\ln b}\ln x = \frac{1}{\ln b}\frac{1}{x}$$

$$\frac{d^2}{dx^2}\log x = \frac{1}{\ln b}\cdot\left(-\frac{1}{x^2}\right) < 0 \quad \text{for} \quad x > 0$$

$\therefore$ for positive x, $\log x$ is concave

(b) Show that $x\log x$ is convex in x for positive x.

Suppose the base of log is b

$\therefore \log_b x = \ln x / \ln b$

$\therefore \frac{d}{dx} x\log x = \frac{d}{dx} \times \frac{\ln x}{\ln b} = \frac{1}{\ln b}[\ln x + 1]$

$\frac{d^2}{dx^2} x\log x = \frac{1}{\ln b} \cdot \frac{1}{x} < 0$ for positive x

$\therefore x\log x$ is convex

(c) Show that $H(p) = -p\log p - (1-p)\log(1-p)$ is concave in p for $0 \le p \le 1$

$H(p) = -p\log p - (1-p)\log(1-p)$

$= -\frac{1}{\ln b}[p\ln p - (1-p)\ln(1-p)]$

$\frac{dH(p)}{dp} = -\frac{1}{\ln b}[\ln p + 1 - \ln(1-p) - 1]$

$\frac{d^2}{dp^2} H(p) = -\frac{1}{\ln b}[\frac{1}{p} + \frac{1}{1-p}]$

$\because 0 \le p \le 1 \quad \therefore \frac{1}{p}, \frac{1}{1-p} \ge 0 \quad \therefore \frac{d^2}{dp^2} H(p) \le 0$

$\therefore H(p)$ is concave in p

for $0 \le p \le 1$

10. Maximum entropy

what is the maximum value of $H(p_1, \cdots p_n) = H(p)$

let u be the uniform distribution on h

$$E_p[-\log u(x)] = E_p[-\log \frac{1}{h}] = \log h$$

$$D(p\|u) = E_p[\log p(x)/u(x)]$$

$$= E_p[\log p(x) - \log u(x)]$$

$$= E_p[-\log u(x)] - E_p[-\log p(x)]$$

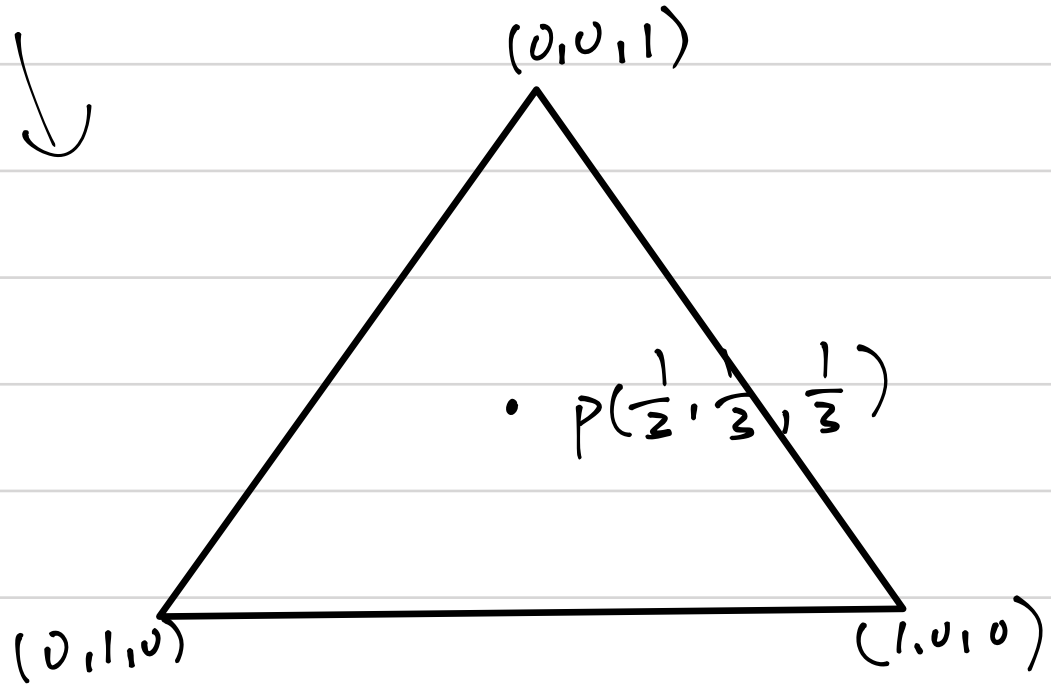
$$= \log h - H(x) \geq 0$$

$$\therefore H(x) \leq \log h$$

$\therefore$ The maximum value of $H(\vec{p})$ is $\log h$, when p is uniform distribution

That is $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$

For $n=3$ $\quad p_1 = p_2 = p_3 = \frac{1}{3}$



11. Drawing with and without replacement

Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement

Let X_i denote the situation for ith draw

$\therefore p(X_i = \text{red}) = r/(r+w+b)$

$p(X_i = \text{white}) = w/(r+w+b)$

$p(X_i = \text{blue}) = b/(r+w+b)$

The entropy we compare is $H(X_1, X_2, \cdots X_k)$

using the chain rule:

$$H(X_1, X_2 \cdots X_k) = \sum_{i=1}^{k} H(X_i \mid X_{i-1}, X_{i-2}, \cdots X_1)$$

① with replacement:

$H(X_i \mid X_{i-1}, X_{i-2} \cdots X_1) = H(X_i)$ (since with replacement, the probability for each draw is same

$H(X_i) = -r/(r+w+b) \log(r/(r+w+b)) - w/(r+w+b)$

$\log(w/(r+w+b)) - b/(r+w+b) \log(b/(r+w+b))$

$= \log(r+w+b) - r/(r+w+b) \log r - w/(r+w+b) \log w - b/(r+w+b) \log b$

② without replacement:

The conditional entropy $H(X_i \mid X_{i-1}, \cdots X_1) \leq H(X_i)$ for each i (since each draw is not independent)

$\therefore$ the entropy of drawing without replacement is smaller

12. Conditional Mutual Information

(a) Show that $X \to Y \to Z$ forms a Markov chain

$I(X;Y|Z) \le I(X;Y)$

Consider $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$

$= I(X;Y) + I(X;Z|Y)$

$\because I(X;Z|Y) = 0$

$\therefore I(X;Z) + I(X;Y|Z) = I(X;Y)$

$I(X;Y|Z) = I(X;Y) - I(X;Y)$

$\because I(X;Y) \ge 0$

$\therefore I(X;Y|Z) \le I(X;Y)$

(b) Not always true

As mentioned in the lecture

Consider X and Y, two independent binary random variables and let $Z = X \oplus Y$

$I(X;Y) = 0 \quad I(X;Y|Z) = 1$

∴ There exists examples that conditions increase mutual information.

13. Find the gap

For $X \to Y \to Z$ $I(X;Z) \le I(Y;Z)$, find the exact gap between these mutual information

$X \to Y \to Z$ also implies $Z \to Y \to X$

Consider $I(Z; X, Y) = I(Z; X) + I(Z; Y|X)$

$= I(Z; Y) + I(Z; X|Y)$

$X \to Y \to Z$ is a markov chain

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

$\therefore I(Z; Y) = I(Z; X) + I(Z; Y|X)$

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

The gap is a conditional mutual information

$I(Y; Z|X)$