COMP2610/6261 Assignment 1 2018 Solutions

1 Entropy and Mutual Information

1. a [2 points] $P(L = A) = \frac{4}{16}$ giving $h(L = A) = 2$
   b [2 points] $P(V = 3) = \frac{4}{16}$ giving $h(V = 3) = 2$
   c [2 points] $P(L = A|V = 1) = \frac{4}{8}$ giving $h(L = A|V = 1) = 1$
   d [2 points] $P(V = 1|L = A) = 1$ giving $h(V = 1|L = A) = 0$, Knowing the letter determines the value hence there is no surprise in discovering the value.
   e [2 points] $H(L) = \frac{11}{4}$, $H(V) = \frac{3}{2}$ and $H(D) \approx 0.811$.

2. [5 points] $I(L;V) = H(V) - H(V|L) = H(V)$ as knowing the letter determines the value of the code meaning $H(V|L) = 0$. Hence $I(L;V) = \frac{3}{2}$

   $I(D;V) = H(V) - H(V|D)$ and we know that $H(V) = \frac{3}{2}$. If $D = \text{decaf}$ then we have the following probability distribution for V

   $P(V = 1|D = \text{decaf}) = \frac{8}{12}$ $P(V = 3|D = \text{decaf}) = \frac{3}{12}$ $P(V = 5|D = \text{decaf}) = \frac{1}{12}$

   Giving $H(V|D = \text{decaf}) \approx 1.189$. Similarly when $D = \text{not decaf}$ then we obtain the following distribution for V

   $P(V = 1|D = \text{not decaf}) = 0$ $P(V = 3|D = \text{not decaf}) = \frac{1}{4}$ $P(V = 5|D = \text{not decaf}) = \frac{3}{4}$

   Giving $H(V|D = \text{not decaf}) \approx 0.811$. As $P(D = \text{decaf}) = \frac{3}{4}$, $H(V|D) \approx 1.094$ meaning $I(V;D) \approx 0.406$.

   For the interpretation, note that $P(V, D, L) = P(V)P(L|V)P(D|L, V) = P(V)P(L|V)P(D|L)$ as a tiles decaf status is determined by its letter. Hence we have a markov chain $V \to L \to D$ so by the information processing theorem $I(V;D) \leq I(V:L)$ as the above calculations have confirmed.

3. a [2 points] $P(D = \text{decaf}|C = \text{cheat}) = \frac{1}{2}$ and $P(D = \text{decaf}|C = \text{not cheat}) = \frac{3}{4}$ giving $H(D|C) \approx \frac{1}{2}(1 + 0.811) = 0.906$
   b [3 points] $H(C, D) = H(C) + H(D|C)$. $H(C) = 1$ giving $H(C, D) \approx 1.905$
   3 [5 points] $P(\text{3 decaf tiles}|\text{cheat}) = (\frac{1}{2})^3$ and $P(\text{3 decaf tiles}|\text{not cheat})(\frac{3}{4})^3$. Hence by Bayes rule

$$\begin{aligned} P(\text{cheat}|\text{3 decaf tiles}) &= \frac{P(\text{3 decaf tiles}|\text{cheat})P(\text{cheat})}{P(\text{3 decaf tiles})} \\ &= \frac{\frac{1}{8}}{\frac{1}{8} + \frac{27}{64}} \\ &= \frac{8}{35} \end{aligned}$$

2 Mutual Information

1. We have

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

Let

$$\begin{aligned}\mathbf{m} = &\left(\frac{p(X=\mathsf{a}|Y=0)+p(X=\mathsf{a}|Y=1)}{2}, \frac{p(X=\mathsf{a}|Y=0)+p(X=\mathsf{a}|Y=1)}{2},\right.\\ &\left.\frac{p(X=\mathsf{a}|Y=0)+p(X=\mathsf{a}|Y=1)}{2}\right)\\ = &\frac{\mathbf{p}+\mathbf{q}}{2}\\ = &\left(\frac{3}{8}, \frac{3}{8}, \frac{1}{4}\right).\end{aligned}$$

Now,

$$\begin{aligned}I(X;Y) &= D_{\mathrm{KL}}(p(X,Y)\|p(X)p(Y))\\ &= \sum_{x,y} -p(x,y)\cdot\log\frac{p(x,y)}{p(x)p(y)}\\ &= \sum_{x,y} -p(x,y)\cdot\log\frac{p(x|y)}{p(x)}\\ &= \sum_{x,y} -p(x|y)\cdot p(y)\cdot\log\frac{p(x|y)}{p(x)}\\ &= \sum_{x} -p(x|y=0)\cdot\log\frac{p(x|y=0)}{p(x)} + \sum_{x} -p(x|y=1)\cdot\log\frac{p(x|y=1)}{p(x)}\\ &= \frac{D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{m})+D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{m})}{2}.\end{aligned}$$

It can be checked that

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{m}) &= -\sum_x p(x)\log\frac{m(x)}{p(x)} \\
&= \sum_x p(x)\log\frac{p(x)}{m(x)} \\
&= \frac{1}{2}\log\frac{4}{3} + \frac{1}{4}\log\frac{2}{3} + \frac{1}{4}\log 1 \\
&= 1 + \frac{1}{4} - \frac{3}{4}\log 3 \\
&\approx 0.0613.
\end{aligned}
$$

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{m}) &= -\sum_x q(x)\log\frac{m(x)}{q(x)} \\
&= \sum_x q(x)\log\frac{q(x)}{m(x)} \\
&= \frac{1}{4}\log\frac{2}{3} + \frac{1}{2}\log\frac{4}{3} + \frac{1}{4}\log 1 \\
&\approx 0.0613.
\end{aligned}
$$

Thus,

$$I(X;Y) \approx 0.0613.$$

2. We have established this above.

3. Let

$$
\begin{aligned}
\mathbf{p}' &= (p(Z=\mathtt{a}|Y=1), p(Z=\mathtt{b}|Y=1), p(Z=\mathtt{c}|Y=1)). \\
\mathbf{q}' &= (p(Z=\mathtt{a}|Y=0), p(Z=\mathtt{b}|Y=0), p(Z=\mathtt{c}|Y=0)). \\
\mathbf{m}' &= \frac{\mathbf{p}'+\mathbf{q}'}{2}.
\end{aligned}
$$

Then, from (ii),

$$I(Z;Y) = \frac{D_{\mathrm{KL}}(\mathbf{p}'\|\mathbf{m}') + D_{\mathrm{KL}}(\mathbf{q}'\|\mathbf{m}')}{2}.$$

From the problem statement,

$$
\begin{aligned}
\mathbf{p}' &= \mathbf{q} \\
\mathbf{q}' &= \mathbf{p}
\end{aligned}
$$

and so

$$\mathbf{m}' = \mathbf{m}.$$

It thus follows that

$$I(Z;Y) = \frac{D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{m}) + D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{m})}{2} = I(X;Y).$$

This is intuitive, because mutual information measures the average reduction in uncertainty in Z when Y is known. Z can be seen as merely "relabelling" the outcomes of X, while sharing the probabilities. Therefore, the uncertainty in both random variables is the same, as is the mutual information with respect to Y.

4. We can construct a Markov chain $Y \to X \to Z$, which would imply that $I(X;Y) \geq I(Z;Y)$. To enforce strict equality, we must ensure that Z somehow discards information from X. For example, let Z be uniformly from $\{\mathtt{a}, \mathtt{b}\}$ if $X = \mathtt{a}$, uniformly from $\{\mathtt{a}, \mathtt{c}\}$ if $X = \mathtt{b}$, and uniformly from $\{\mathtt{b}, \mathtt{c}\}$ if $X = \mathtt{c}$. Then, we have that

$$
\begin{aligned}
p(Z = \mathtt{a}|Y = y) &= \frac{p(X = \mathtt{a}|Y = y) + p(Z = \mathtt{b}|Y = y)}{2} \\
p(Z = \mathtt{b}|Y = y) &= \frac{p(X = \mathtt{a}|Y = y) + p(Z = \mathtt{c}|Y = y)}{2} \\
p(Z = \mathtt{c}|Y = y) &= \frac{p(X = \mathtt{b}|Y = y) + p(Z = \mathtt{c}|Y = y)}{2},
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathbf{p}' &= (3/8, 3/8, 1/4) \\
\mathbf{q}' &= (3/8, 1/4, 3/8).
\end{aligned}
$$

It can now be checked, using the formula from part (ii), that

$$I(Z;Y) = 0.0182 < I(X;Y).$$

5. We can construct a Markov chain $Y \to Z \to X$, where now X is a noisy version of Z. For example, let Z have conditional probability vectors

$$\mathbf{p}' = (1/2, 1/2, 0)$$

$$\mathbf{q}' = (1/2, 0, 1/2)$$

and say that X is uniform from $\{\mathtt{a}, \mathtt{b}\}$ if $Z = \mathtt{a}$, uniform from $\{\mathtt{a}, \mathtt{c}\}$ if $Z = \mathtt{b}$, and uniform from $\{\mathtt{b}, \mathtt{c}\}$ if $Z = \mathtt{c}$. It may be verified that this results in the same $\mathbf{p}, \mathbf{q}$ for X as in part (i); it may further be seen that $I(Z;Y) = +\infty > I(X;Y)$, with the infinite mutual information owing to the KL divergence blowing up due to the presence of zero entries in $\mathbf{p}', \mathbf{q}'$.

3 Uniform Length, Lossy Coding

Denote the albums by their first letters and let $\mathcal{A}=\{A, B, C, D\}$.

1. Bits required to uniformly code:

   (a) Two bits ($= \log_2 4$). Example uniform code: $C = \{00, 01, 10, 11\}$.

   (b) The album Alina has 5 tracks so $\lceil \log_2 5 \rceil = 3$ bits are required. Example code: $C = \{000, 001, 010, 011, 100\}$.

(c) There are 44 tracks in total so $\lceil \log_2 44 \rceil = 6$ bits are required.

2. The raw bit content is $\log_2 44 \approx 5.46$.

3. Let A denote ensemble of album chosen when tracks across all albums are picked uniformly at random.

   (a) $\mathcal{A}_A = \{A, B, C, D\}$ and $p = \{5/44, 12/44, 15/44, 12/44\}$.

   (b) Raw bit content $= \log_2 |\mathcal{A}_{A^4}| = \log_2 4^4 = 8$.

   (c) Since A^4 has 256 elements we need a δ equal to the smallest probability for element of A^4. As album A has smallest probability, so too will $AAAA \in \mathcal{A}_A$. Thus choose $\delta = P(AAAA) = (5/44)^4 \approx 0.0002$.

   (d) $H_\delta(A^4)$ will be zero when S_δ contains only one element so if δ is set so that there are two elements in S_δ we are done. Since Coieda has the highest probability the sequence $CCCC \in \mathcal{A}_{A^4}$ will have highest probability $P(CCCC) = (15/44)^4$. The next largest probability is any sequence with three Cs and either B or D in it. $P(BCCC) = (12/44)(15/44)^3$ so set $\delta = (12/44)(15/44)^3 + (15/44)^4 \approx 0.024$.

4. (a) $H(A) = \frac{1}{44}\left(5\log_2(5/44) + 2 \times 12\log_2(12/44) + 15\log_2(15/44)\right) \approx 1.91$

   (b) Items in typical set have $P(a)$

   (c) The probability of sequences of length $N = 100$ in $T_{N\beta}$ for $\beta = 0.1$ is no more than $2^{-100(H(A)-0.1)} \approx 3.26 \times 10^{-55}$ and no less than $2^{-100(H(A)+0.1)} \approx 3.11 \times 10^{-61}$. This means $\frac{1}{3.26\times10^{-55}} \approx 3.1 \times 10^{54} \leq |T_{N\beta}| \leq \frac{1}{3.11\times10^{-61}} \approx 3.2 \times 10^{60}$ and so there are approximately 10^{57} elements.

   (d) No, because that rate is below the entropy of 1.91 so, by the source coding theorem, for large blocks a rate of 1.5 bits per title is not possible.

Q4 (c) Complete part (c) by setting $c = \frac{\tau}{2}$ ..

We need to compute $\mathbb{E}\, l_\tau(X-c) = \int_{-\infty}^{\infty} l_\tau(x-c) f(x)\,dx$ where $f(\cdot)$ is the density of X.

Observe that
$$l_\tau(x-c) = \begin{cases} \tau(x-c) & x \geq c \\ (\tau-1)(x-c) & x \leq c. \end{cases}$$

Thus
$$\mathbb{E}\, l_\tau(X-c) = \int_{-\infty}^{c} (\tau-1)(x-c) f(x)\,dx + \int_c^{\infty} \tau(x-c) f(x)\,dx$$

$$= (\tau-1)\int_{-\infty}^{c} x f(x)\,dx - (\tau-1)c\int_{-\infty}^{c} f(x)\,dx + \tau\int_c^{\infty} x f(x)\,dx - \tau c\int_c^{\infty} f(x)\,dx$$

$$= (\tau-1)E_{-\infty}^{c} - (\tau-1)cF(c) + \tau E_c^{\infty} - \tau c(1-F(c)) \qquad (*)$$

where $F(c)$ is the cumulative distribution of X, and $E_a^b := \int_a^b x f(x)\,dx$.

Observe that by Leibnitz's rule, (or just fundamental theorem of calculus)
$$\frac{\partial}{\partial c} E_{-\infty}^{c} = \frac{\partial}{\partial c}\int_{-\infty}^{c} \varphi(x)\,dx = \varphi(c)$$
where $\varphi(x) = x f(x)$

and
$$\frac{\partial}{\partial c} E_c^{\infty} = \frac{\partial}{\partial c}\int_c^{\infty} \varphi(x)\,dx = -\varphi(c).$$

Thus from $(*)$ we have
$$\frac{\partial}{\partial c}\mathbb{E}\, l_\tau(X-c) = (\tau-1)cf(c) - (\tau-1)cf(c) - (\tau-1)F(c) - \tau c f(c) + \tau c f(c) - \tau(1-F(c))$$
$$= cf(c)\underbrace{\left[(\tau-1) - (\tau-1) - \tau + \tau\right]}_{=0} - (\tau-1)F(c) - \tau(1-F(c))$$

Set to zero:
$$-\tau F(c) + F(c) - \tau + \tau F(c) = 0$$
$$\Rightarrow \quad F(c) = \tau \quad \Rightarrow \quad c = q_\tau$$

4 (b).

$$
\begin{aligned}
|\mu - m| &= \mathbb{E}(X - m) \\
&\leq \mathbb{E}|X - m| && \text{(Jensen)} \\
&\leq \mathbb{E}|X - \mu| && \text{(from 4a)} \\
&\leq \sqrt{\mathbb{E}(X-\mu)^2} && \text{(Jensen)} \\
&= \sigma && \text{(Def}^{n}\text{ of } \sigma)
\end{aligned}
$$

4(d). $l_\alpha(x) = 0$ only if $x = 0$.

Thus $\mathbb{E}\, l_\alpha(X - c) = 0$ only if $X \equiv c$.

Suppose $Q_\alpha(X) = \min_{c \in \mathbb{R}} \mathbb{E}\, l_\alpha(X-c) = q \neq 0$

Consider $Y = \beta X$, for some $\beta > 0$.

$Q_\alpha(Y) = Q_\alpha(\beta X) = \min_c \mathbb{E}\, l_\alpha(\beta X - c)$

Now $l_\alpha(\beta x) = \beta\, l_\alpha(x)$

The minimizing c value will differ, but we see that

$$Q_\alpha(\beta X) = \beta\, Q_\alpha(X)$$

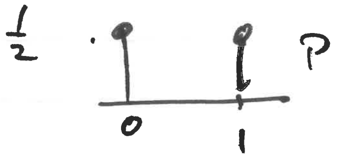
Compare $\sigma(\beta X) = \beta\, \sigma(X)$.

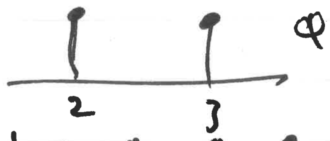
σ^2 does not always exist. But every random variable has quantiles.
[There is a subtlety when the distribution of X has atomic components, but the key point is that Q_α always exists.]

$Q_\alpha(X)$ corresponds to $CVar_\alpha(X - E(X))$
the conditional value at risk of $X - E(X)$

– see [R. Tyrell Rockafellar & Stan Uryasev, "The Fundamental Risk Quadrangle in Risk Management, Optimization and Statistical Estimation", Surveys in Operations Research and Management Science 18 (1-2), 33-53 (2013)]

4 e. Consider two discrete distributions





$H = \log_2 2 = 1$

equal mixture is

$H = 4 \cdot \frac{1}{4} \log_2 4$
$= 2$

So $H(P) = H(Q) = 1$
$H(\frac{1}{2}P + \frac{1}{2}Q) = 2$.

5). Assume w.l.o.g. $\mu = E(X) = 0$.
We assume $X \in [a, b]$ (as assumed in ~~part~~ (d)).

Can write $X = \alpha b + (1-\alpha) a$ where
$\alpha = \frac{(X-a)}{b-a}$.

Since $g(y) = e^{ty}$ is convex

$$e^{tX} \leq \alpha e^{tb} + (1-\alpha) e^{ta}$$
$$= \frac{X-a}{b-a} e^{tb} + \frac{b-X}{b-a} e^{ta}.$$

Take expectations of both sides to obtain

$$E(e^{tX}) \leq -\frac{a}{b-a} e^{tb} + \frac{b}{b-a} e^{ta} = e^{g(\underbrace{t(b-a)}_{u})}$$

where $u = t(b-a)$ & $g(u) = -\gamma u + \log(1-\gamma+\gamma e^{u})$;
$\gamma = \frac{-a}{b-a}$

b) Since $E(X) = 0$, $a \leq 0 \leq b$.

$$\frac{\partial^2}{\partial u^2} g(u) = \frac{\gamma e^{u}}{1-\gamma+\gamma e^{u}} - \frac{\gamma^2 (e^{u})^2}{(1-\gamma+\gamma e^{u})^2}$$

$$= \frac{\gamma e^{u}(1-\gamma+\gamma e^{u}) - \gamma^2 (e^{u})^2}{(1-\gamma+\gamma e^{u})^2}$$

Denominator is always non-negative.

Numerator can be expanded as

$$\gamma e^{u} - \gamma^2 e^{u} + \cancel{\gamma^2 (e^{u})^2 - \gamma^2 (e^{u})^2}$$
$$= \gamma e^{u}(1-\gamma).$$ But $\gamma = \frac{-a}{b-a} \leq \frac{-a}{a} = 1$

Thus $\frac{\partial^2}{\partial u^2} g(u) \geq 0$ & g is cvx.

Observe that $\frac{\partial^2}{\partial u^2} g(u) = \dfrac{\gamma e^u[(1-\gamma+\gamma e^u) - \gamma e^u]}{(1-\gamma+\gamma e^u)^2}$

$$= \frac{\gamma e^u(1-\gamma)}{(1-\gamma+\gamma e^u)^2} \qquad \gamma \in [0,1]$$

One can check numerically that $\frac{\partial^2}{\partial u^2} g(u) \leq \frac{1}{4}$ $\forall u > 0$.

By Taylor's Theorem there is $\xi \in (0,u)$ such that

$$g(u) = g(0) + u g'(0) + \frac{u^2}{2} g''(\xi) = \frac{u^2}{2} g''(\xi)$$

since $g(0) = g'(0) = 0$. Hence

$$g(u) \leq \frac{u^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Thus $\mathbb{E} e^{tX} \leq e^{g(u)} \leq e^{t^2(b-a)^2/8}$.

---

c) $P(X > \varepsilon) = P(e^X > e^\varepsilon) = P(e^{tX} > e^{t\varepsilon})$ for any $t \geq 0$

But $P(e^{tX} > e^{t\varepsilon}) \leq e^{-t\varepsilon} \mathbb{E}(e^{tX})$ by Markov's ≠.

$\left(P(Y > s) \leq \frac{1}{s} E(Y) \right)$

This holds for any $t > 0$, & thus

$$P(X > \varepsilon) \leq \inf_{t \geq 0} e^{-t\varepsilon} \mathbb{E}(e^{tX}).$$

d) Assume w.l.o.g. that $\mu = E(X_i) = 0$.

$$P(|\bar{X}_n| \geq \varepsilon) = P(\bar{X}_n \geq \varepsilon) + P(\bar{X}_n \leq -\varepsilon)$$

$$= P(\bar{X}_n \geq \varepsilon) + P(-\bar{X}_n \geq \varepsilon)$$

Now using earlier results, we have

$$P(\bar{X}_n \geq \varepsilon) = P\left(\sum_{i=1}^{n} X_i \geq n\varepsilon\right) = P\left(e^{\sum_{i=1}^{n} X_i} \geq e^{n\varepsilon}\right)$$

$$= P\left(e^{t\sum_{i=1}^{n} X_i} \geq e^{tn\varepsilon}\right)$$

$$\leq e^{-tn\varepsilon} E\left(e^{t\sum_{i=1}^{n} X_i}\right)$$

$$= e^{-tn\varepsilon} E\left(\prod_{i=1}^{n} e^{tX_i}\right)$$

$$= e^{-tn\varepsilon} \prod_{i=1}^{n} E(e^{tX_i}) \qquad \text{(by independence)}$$

$$= e^{-tn\varepsilon} \left(E(e^{tX_i})\right)^n \qquad (\text{for any } i \in [n])$$

By part (b) $\quad E(e^{tX_i}) \leq e^{t^2(b-a)^2/8}$

Then $\quad P(\bar{X}_n \geq \varepsilon) \leq e^{-tn\varepsilon} e^{t^2 n(b-a)^2/8} \qquad (*)$

This holds for all t, so we can optimize over t

$$\frac{\partial}{\partial t}\, \cancel{e^{-tn\varepsilon}\, e^{t^2 n(b-a)^2/8}}$$

$$= \cancel{-n\varepsilon\, e^{-tn\varepsilon}\, e^{t^2n(b-a)^2/8} + e^{-tn\varepsilon} \cdot e^{t^2}}$$

Take logs $-tn\varepsilon + t^2 n(b-a)^2/8$
(exp is monotone).

$$\frac{\partial}{\partial t}\left(-tn\varepsilon + t^2 n(b-a)^2/8\right)$$

$$= -n\varepsilon + 2tn(b-a)^2/8$$

Set to zero & solve for t

$$\frac{2tn(b-a)^2}{8} = n\varepsilon$$

$$\Rightarrow \quad t = \frac{4\varepsilon}{(b-a)^2}.$$

Substitute this value of t, so (*) becomes

$$P(\bar{X}_n \geq \varepsilon) \leq e^{-\frac{4\varepsilon}{(b-a)^2}\cdot n\varepsilon}\, e^{+\frac{16\varepsilon^2}{(b-a)^4}\cdot\frac{(b-a)^2 n}{8}}$$

$$= \cancel{e^{-\frac{4\varepsilon^2 n}{(b-a)^2}} \cdot e^{-2\varepsilon^2}}$$

$$e^{-\frac{4n\varepsilon^2}{(b-a)^2}} \cdot e^{+\frac{2\varepsilon^2 n}{(b-a)^2}}$$

$$= e^{-2n\varepsilon^2/(b-a)^2}.$$

A similar argument holds for $P(\bar{X}_n \leq \varepsilon)$

So we obtain the result by $P(A \text{ or } B) = P(A) + P(B)$ given A, B mutually exclusive

5e). Bernoulli $\Rightarrow (b-a) = 1$.