COMP2610/COMP6261
Tutorial 5 Sample Solutions

Tutorial 5: Probabilistic inequalities and Mutual Information

Young Lee and Bob Williamson
Tutors: Debashish Chakraborty and Zakaria Mhammedi

Week 5 (21st - 25th August), Semester 2, 2017

1. Consider a discrete variable X taking on values from the set $\mathcal{X}$. Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$. Denote the vector of probabilities by $\mathbf{p}$. We saw in lectures that the entropy of X satisfies:

$$H(X) \leq \log|\mathcal{X}|,$$

with equality if and only if $p_i = \dfrac{1}{|\mathcal{X}|}$ for all i, i.e. $\mathbf{p}$ is uniform. Prove the above statement using Gibbs' inequality, which says

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{p_i}{q_i} \geq 0$$

for any probability distributions $\mathbf{p}, \mathbf{q}$ over $|\mathcal{X}|$ outcomes, with equality if and only if $\mathbf{p} = \mathbf{q}$.

Solution.

Gibb's inequality tells us that for any two probability vectors $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$ and $\mathbf{q} = (q_1, \ldots, q_{|\mathcal{X}|})$:

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{q_i} \geq 0$$

with equality if and only if $\mathbf{p} = \mathbf{q}$. If we take $\mathbf{q}$ to be the vector representing the uniform distribution $q_1 = \ldots = q_{|\mathcal{X}|} = \dfrac{1}{|\mathcal{X}|}$, then we get

$$0 \leq \sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{\frac{1}{|\mathcal{X}|}} = \sum_{i=1}^{|\mathcal{X}|} p_i \log p_i + \sum_{i=1}^{|\mathcal{X}|} p_i \log|\mathcal{X}| = -H(\mathbf{p}) + \log|\mathcal{X}|$$

with equality if and only if $\mathbf{p}$ is the uniform distribution. Moving $H(\mathbf{p})$ to the other side gives the inequality.

2. Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$\begin{aligned} H(X, g(X)) &\stackrel{(a)}{=} H(X) + H(g(X)|X) \\ &\stackrel{(b)}{=} H(X); \\ H(X, g(X)) &\stackrel{(c)}{=} H(g(X)) + H(X|g(X)) \\ &\stackrel{(d)}{\geq} H(g(X)). \end{aligned}$$

Thus $H(g(X)) \leq H(X)$.

Solution.

(a) This is using the chain rule of entropy,
i.e. $H(X, Y) = H(X) + H(Y \mid X)$ where $Y = g(X)$

(b) Given X, we can determine $g(X)$ since it is fixed, being a function of X. This means no uncertainty remains about $g(X)$ when X is given. Thus, $H(g(X) \mid X) = 0$ since $\sum_x p(x)p(g(X) \mid X = x) = 0$.

(c) This is also using the chain rule of entropy,
i.e. $H(X, Y) = H(Y) + H(X \mid Y)$ where $Y = g(X)$

(d) In this case, $H(X \mid g(X)) \geq 0$ since the conditional entropy of a discrete random variable is non-negative. If $g(X)$ has one-to-one mapping with X, then $H(X, g(X)) \geq H(g(X))$.

Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

3. Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if their joint probability distribution can be written as:

$$p(X, Y, Z) = p(X) \cdot p(Y|X) \cdot p(Z|Y)$$

(a) Suppose (X, Y, Z) forms a Markov chain. Is it possible for $I(X; Y) = I(X; Z)$? If yes, give an example of X, Y, Z where this happens. If no, explain why not.

(b) Suppose (X, Y, Z) does *not* form a Markov chain. Is it possible for $I(X; Y) \geq I(X; Z)$? If yes, give an example of X, Y, Z where this happens. If no, explain why not.

Solution.

(a) Yes; pick $Z = Y$.

Reason: The data processing inequality guarantees $I(X; Y) \geq I(X; Z)$. Here we want to verify that equality is possible. If we look at the proof of the data processing inequality, we just need to find a Z where $I(X; Y|Z) = 0$.

For $Z = Y$, intuitively, conditioning on Z, the reduction in uncertainty in X when we know Y is zero, because Z already tells us everything that Y can. Formally,$I(X; Y) = I(X; Z)$ because the random variables Y and Z have the same distribution. Note: to formally check that Z is conditionally independent of X given Y, we can check $p(Z = z, X = x|Y = y) = p(Z = z|Y = y) \cdot p(X = x|Y = y)$ for all possible x, y, z. The reason is that the left and right hand sides are zero when $y \neq z$; and when $y = z$, they both equal $p(X = x|Y = y)$ as $p(Z = z|X = x, Y = y) = 1$ in this case.

(b) Yes; pick X, Z independent, and let $Y = X + Z$ (assuming the outcomes are numeric).

Reason: Z is not conditionally independent of X given Y ; intuitively, knowing $X + Z$ and X tells us what Z is. So (X, Y, Z) does not form a Markov chain. However, since X, Z are independent, $I(X; Z) = 0$. Since mutual information is non-negative, $I(X; Y) \geq 0 = I(X; Z)$.

4. If $X \to Y \to Z$, then show that

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

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

   Proof in lecture 9

5. A coin is known to land heads with probability $\frac{1}{5}$. The coin is flipped N times for some even integer N.

   (a) Using Markov's inequality, provide a bound on the probability of observing $\frac{N}{2}$ or more heads.

   (b) Using Chebyshev's inequality, provide a bound on the probability of observing $\frac{N}{2}$ or more heads. Express your answer in terms of N.

   (c) For $N \in \{2, 4, \ldots, 20\}$, in a single plot, show the bounds from part (a) and (b), as well as the *exact* probability of observing $\frac{N}{2}$ or more heads.

   Solution.

   $X_1, \ldots, X_N$ represents N flips, where, independent bernoulli random variable, $X_i = 1$ represents observing head from a coin flip and $X_i = 0$ represents observing tail. Suppose $\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. So, the probability of observing $\frac{N}{2}$ heads can be expressed as $p(\hat{X}_N \ge \frac{1}{2})$ and $p(X_i = 1) = \frac{1}{5}$ for each i.

   (a) Using Markov's Inequality,

$$p(\hat{X}_N \ge \frac{1}{2}) \le \frac{E[\hat{X}_N]}{\frac{N}{2}}$$
$$= \frac{\frac{\sum_{i=1}^{N} E[X_i]}{N}}{\frac{1}{2}} = \frac{\frac{1}{5}}{\frac{1}{2}} = \frac{2}{5}$$

$$\therefore p(\hat{X}_N \ge \frac{N}{2}) \le \frac{2}{5}$$

   (b) We need to calculate the variance of the bernoulli random variable: $Var(X) = p(1-p)$

$$\therefore Var[X_i] = (\frac{1}{5})(1 - \frac{1}{5}) = \frac{4}{25}$$

   Using the definition of variance and its properties,

$$Var(\hat{X}_N) = Var[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{\sum_{i=1}^{N} Var[X_i]}{N^2} = \frac{N(\frac{4}{25})}{N^2} = \frac{4}{25N}$$

   Using Chebyshev's inequality,

$$p(\mid \hat{X}_N - E[\hat{X}_N] \mid \ge \lambda) \le \frac{Var(\hat{X}_N)}{\lambda^2}$$
$$p(\mid \hat{X}_N - \frac{1}{5} \mid \ge \frac{3}{10}) \le \frac{\frac{4}{25N}}{(\frac{3}{10})^2}$$
$$p(\hat{X}_N \ge \frac{1}{2}) \le \frac{16}{9N}$$

   (c) The exact probability of a k heads is given by the binomial distribution:

$$P(X = k) = \binom{N}{k}(\frac{1}{5})^k(\frac{4}{5})^{N-k}$$

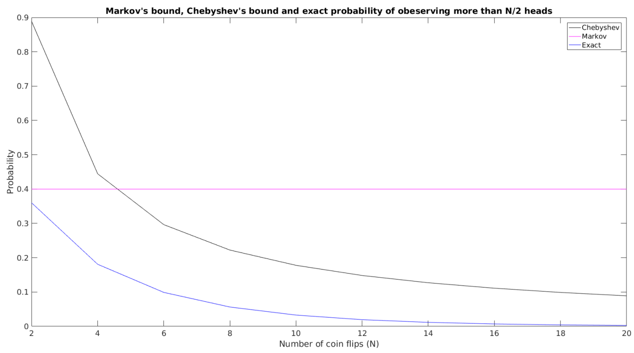
So, the probability of seeing N/2 or more heads is

$$
\begin{aligned}
P(X \geq N/2) &= \sum_{k=N/2}^{N} P(X = k) \\
&= \sum_{k=N/2}^{N} \binom{N}{k} (\frac{1}{5})^k (\frac{4}{5})^{N-k}
\end{aligned}
$$

Another way to calculate the exact probability

$$
p(\hat{X}_N \geq \frac{1}{2}) = 1 - p(\hat{X}_N < \frac{1}{2})
$$

This can be done in Matlab using (1-binocdf(floor(0.5.*n-0.5), n, 0.2))
Here floor(0.5.*n-0.5) simply brings the value of n to an integer less than $n/2$ for each value of n. For example, a value of n=10 would lead floor(0.5.*n-0.5) value of 4, which is what we want.



The code for the plot above is included below:

```matlab
n = 2:2:20;

% Markov Inequality
y_m = 2/5;

% Chebyshev Inequality
y_c = 16 ./(9 .*n);

% Exact Probabilities
y_e = 1-binocdf(floor(0.5.*n-0.5), n, 0.2);

plot(n,y_c,'k')
hold on;
plot([2 20],[y_m y_m], 'm-')
hold on;
plot(n, y_e,'b')
hold on;
set(gca,'fontsize', 14)

title('Markov''s bound, Chebyshev''s bound and exact probability of obeserving more than N/2 heads')
ylabel('Probability')
xlabel('Number of coin flips (N)')
legend('Chebyshev', 'Markov','Exact');
```