ASSIGNMENT 1

Suggested Solutions

Contributors:
Raymond W. Yeung
Ning Cai
Terence H. Chan
Siu Wai Ho
Shenghao Yang
Zhixue Zhang

CHAPTER 2

1. Let X and Y be random variables with alphabets $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4, 5\}$ and joint distribution $p(x, y)$ given by

$$\frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 1 & 1 \\ 0 & 3 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 3 \end{bmatrix}.$$

Determine $H(X), H(Y), H(X|Y), H(Y|X)$, and $I(X;Y)$.

Solution:

$H(X) = H(Y) = \log 5$.
$H(X,Y) = 2\log 5 - \frac{8}{25}\log 2 - \frac{6}{25}\log 3$.
$H(X|Y) = H(X,Y) - H(Y) = \log 5 - \frac{8}{25}\log 2 - \frac{6}{25}\log 3$.
$H(Y|X) = H(X,Y) - H(X) = \log 5 - \frac{8}{25}\log 2 - \frac{6}{25}\log 3$.
$I(X;Y) = H(X) + H(Y) - H(X,Y) = \frac{8}{25}\log 2 + \frac{6}{25}\log 3$.

2. Prove Propositions 2.8, 2.9, ~~2.10~~, 2.19, 2.21, and 2.22.

   Solution:

   Proof of Proposition 2.8

   We will first prove the 'only if' part by induction on n. The claim is true for $n = 3$. Assume it is true for all $n \le m$, where $m \ge 3$, and consider the Markov chain $X_1 \to X_2 \to \cdots \to X_m \to X_{m+1}$. Then by (2.15),

$$\begin{aligned} & p(x_2)p(x_3)\cdots p(x_m)p(x_1, x_2, \cdots, x_m, x_{m+1}) \\ = \;& p(x_1, x_2)\cdots p(x_{m-1}, x_m)p(x_m, x_{m+1}). \end{aligned}$$

   Summing over all x_{m+1}, we have

$$\begin{aligned} & p(x_2)\cdots p(x_{m-1})p(x_m)p(x_1, x_2, \cdots, x_m) \\ = \;& p(x_1, x_2)\cdots p(x_{m-1}, x_m)p(x_m). \end{aligned}$$

   If $p(x_m) > 0$, then cancel $p(x_m)$ on both sides to obtain

$$p(x_2)\cdots p(x_{m-1})p(x_1, x_2, \cdots, x_m) = p(x_1, x_2)\cdots p(x_{m-1}, x_m). \tag{A2.1}$$

   Otherwise, $p(x_1, x_2, \cdots, x_m) \le p(x_m) = 0$ implies $p(x_1, x_2, \cdots, x_m) = 0$. Similarly, we see that $p(x_{m-1}, x_m) = 0$. Thus (A2.1) continues to be valid for $p(x_m) = 0$. By Definition 2.4, we have $X_1 \to X_2 \to \cdots \to X_m$, and so by the induction hypothesis,

$$\begin{aligned} & X_1 \to X_2 \to X_3 \\ & (X_1, X_2) \to X_3 \to X_4 \\ & \qquad\qquad \vdots \\ & (X_1, X_2, \cdots, X_{m-2}) \to X_{m-1} \to X_m. \end{aligned}$$

   It remains to show that

$$(X_1, \cdots, X_{m-2}, X_{m-1}) \to X_m \to X_{m+1}. \tag{A2.2}$$

   Toward this end, we write

$$\begin{aligned} & p(x_1, \cdots, x_m, x_{m+1}) \\ = \;& \begin{cases} \frac{p(x_1,x_2)\cdots p(x_{m-1},x_m)p(x_m,x_{m+1})}{p(x_2)\cdots p(x_m)} & \text{if } p(x_2), \cdots, p(x_m) > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Define

$$f(x_1, \cdots, x_m) = \begin{cases} \frac{p(x_1,x_2)\cdots p(x_{m-1},x_m)}{p(x_2)\cdots p(x_m)} & \text{if } p(x_2), \cdots, p(x_m) > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(x_m, x_{m+1}) = p(x_m, x_{m+1}).$$

If $p(x_m) > 0$ and $p(x_2), \cdots, p(x_{m-1}) > 0$, then

$$p(x_1, \cdots, x_m, x_{m+1}) = f(x_1, \cdots, x_m) g(x_m, x_{m+1}). \tag{A2.3}$$

If $p(x_m) > 0$ and $p(x_i) = 0$ for some $2 \leq i \leq m-1$, then $p(x_1, \cdots, x_m, x_{m+1}) = 0$ and $f(x_1, \cdots, x_m) = 0$, so that (A2.3) again holds. Thus, (A2.3) holds whenever $p(x_m) > 0$. By Proposition 2.5, the Markov chain in (A2.2) is established, completing the proof for the 'only if' part.

We now prove the 'if' part. Assume that

$$\begin{aligned} & X_1 \to X_2 \to X_3 \\ & (X_1, X_2) \to X_3 \to X_4 \\ & \qquad \vdots \\ & (X_1, X_2, \cdots, X_{m-2}) \to X_{m-1} \to X_m. \end{aligned}$$

If $p(x_2), p(x_3), \cdots, p(x_{m-1}) > 0$, then

$$\begin{aligned} & p(x_1, x_2, \cdots, x_m) \\ = \;& p(x_1, x_2, \cdots, x_{m-1}) p(x_m | x_{m-1}) \\ = \;& p(x_1, x_2, \cdots, x_{m-2}) p(x_{m-1} | x_{m-2}) p(x_m | x_{m-1}) \\ \vdots \;& \\ = \;& p(x_1, x_2) p(x_3 | x_2) \cdots p(x_m | x_{m-1}). \end{aligned}$$

On the other hand, if $p(x_i) = 0$ for some $2 \leq i \leq m-1$, then $p(x_1, x_2, \cdots, x_m) \leq p(x_i) = 0$, which implies $p(x_1, x_2, \cdots, x_m) = 0$. Thus we have shown that $X_1 \to X_2 \to \cdots \to X_m$, proving the 'if' part of the proposition. Hence, the proposition is proven.

Proof of Proposition 2.9

It suffices to show that

$$p(x_1, \cdots, x_n) = f_1(x_1, x_2) \cdots f_{n-1}(x_{n-1}, x_n) \tag{A2.4}$$

if $p(x_2), \cdots, p(x_{n-1}) > 0$ iff

$$p(x_1, \cdots, x_n) = \begin{cases} \frac{p(x_1,x_2)\cdots p(x_{n-1},x_n)}{p(x_2)\cdots p(x_{n-1})} & \text{if } p(x_2), \cdots, p(x_{n-1}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The 'if' part is trivial and its proof is omitted. We now prove the 'only if' part. Define for $1 \leq i \leq n$,

$$\begin{aligned} Q(i) &= \sum_{x_1} \cdots \sum_{x_{i-1}} f_1(x_1, x_2) \cdots f_{i-1}(x_{i-1}, x_i) \\ S(i) &= \sum_{x_{i+1}} \cdots \sum_{x_n} f_i(x_i, x_{i+1}) \cdots f_{n-1}(x_{n-1}, x_n) \end{aligned}$$

with the convention that $Q(1) = S(n) = 1$. Then by summing over all x_j for $j \neq i-1, i$ in (A2.4), it is not difficult to show that

$$p(x_{i-1}, x_i) = f_{i-1}(x_{i-1}, x_i) Q(i-1) S(i) \tag{A2.5}$$

for $1 \leq i \leq n$. Summing over all x_{i-1} in the above, we can further obtain

$$p(x_i) = Q(i) S(i). \tag{A2.6}$$

Hence, by using the expressions for $p(x_{i-1}, x_i)$ and $p(x_i)$, and cancelling the corresponding terms, we obtain

$$\begin{aligned} \frac{p(x_1, x_2) \cdots p(x_{n-1}, x_n)}{p(x_2) \cdots p(x_{n-1})} &= f_1(x_1, x_2) \cdots f_{n-1}(x_{n-1}, x_n) Q(1) S(n) \\ &= f_1(x_1, x_2) \cdots f_{n-1}(x_{n-1}, x_n) \cdot 1 \cdot 1 \\ &= p(x_1, \cdots, x_n). \end{aligned}$$

No need to prove Proposition 2.10 for Assignment 1.

This Proof is included for your self study only.

Proof of Proposition 2.10

Let i_j be the largest element in α_j, $1 \leq j \leq m$, and $i_0 = 0$. Define $\gamma_j = \{i_{j-1}+1, \cdots, i_j\}$, so that $\alpha_j \subset \gamma_j$, and let $\beta_j = \gamma_j \setminus \alpha_j$. Consider a Markov chain $X_1 \to X_2 \to \cdots \to X_n$. By Proposition 2.9,

$$p(x_1, \cdots, x_n) = f_1(x_1, x_2) \cdots f_{n-1}(x_{n-1}, x_n) \tag{A2.7}$$

for all $x_1, x_2, \cdots, x_n$ such that $p(x_2), \cdots, p(x_{n-1}) > 0$. By defining

$$f_k^*(x_k, x_{k+1}) = \begin{cases} f_k(x_k, x_{k+1}) & \text{if } p(x_{k+1}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

(Con't) No need to prove Proposition 2.10 for assignment 1.   This Proof is included for your self study only.

for $1 \leq k \leq n-1$, we have

$$p(x_1, \cdots, x_n) = f_1^*(x_1, x_2) \cdots f_{n-1}^*(x_{n-1}, x_n) \tag{A2.8}$$

for all $x_1, \cdots, x_n$. Note that $f_k^*(x_k, x_{k+1})$ is well-defined because if $p(x_{k+1}) > 0$, then $p(x_1, \cdots, x_n) > 0$ for some $x_1, \cdots, x_k, x_{k+1}, \cdots, x_n$, which implies that $p(x_2), \cdots, p(x_{n-1}) > 0$. For notational convenience, we will let X_0 be a constant and define the function

$$f_0^*(x_0, x_1) = \begin{cases} 1 & \text{if } p(x_1) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Denote $(x_l, l \in \gamma_j)$ by x_{γ_j}. For $0 \leq j \leq m-1$, let

$$g_j(x_{i_j}, x_{\gamma_{j+1}}) = f_j^*(x_{i_j}, x_{i_j+1}) f_{j+1}^*(x_{i_j+1}, x_{i_j+2}) \cdots f_{i_{j+1}-1}^*(x_{i_{j+1}-1}, x_{i_{j+1}}),$$

We also let

$$G(x_{i_m}, \cdots, x_n) = f_{i_m}^*(x_{i_m}, x_{i_{m+1}}) \cdots f_{n-1}^*(x_{n-1}, x_n).$$

Then (A2.8) can be written as

$$p(x_1, \cdots, x_n) = \left[\prod_{j=0}^{m-1} g_j(x_{i_j}, x_{\gamma_{j+1}}) \right] G(x_{i_m}, \cdots, x_n). \tag{A2.9}$$

Denote $\prod_{l \in A} \mathcal{X}_l$ by $\mathcal{X}_A$ and fix $x_{\alpha_j}, 1 \leq j \leq m$. Summing over all vectors $(x_1', \cdots, x_n')$ such that $x_{\alpha_j}' = x_{\alpha_j}$ for $1 \leq j \leq m$ in (A2.9), we have

$$p(x_{\alpha_1}, \cdots, x_{\alpha_m}) = \left[\prod_{j=0}^{m-1} \sum g_j(x_{i_j}, x_{\gamma_{j+1}}) \right] \sum G(x_{i_m}, \cdots, x_n),$$

where the summation inside the square brackets is taken over all the vectors in $\mathcal{X}_{\beta_j}$, while the other summation is taken over all the vectors in $\prod_{l=i_m+1}^{n} \mathcal{X}_l$. For $j = 0$, the summation $\sum g_j(x_{i_j}, x_{\gamma_{j+1}}) = \sum g_0(x_0, x_{\gamma_1})$ depends only on x_{α_1} because x_0 is a constant, and hence we can write it as $f_0'(x_{\alpha_1})$. For $1 \leq j \leq m-1$, $\sum g_j(x_{i_j}, x_{\gamma_{j+1}})$ depends only on X_{i_j} and x_{α_j}, and hence we can write it as $f_j'(x_{\alpha_j}, x_{\alpha_{j+1}})$. Finally, $\sum G(x_{i_m}, \cdots, x_n)$ depends only on X_{i_m}, and hence we can write it as $G'(x_{\alpha_m})$. Therefore, we have

$$p(x_{\alpha_1}, \cdots, x_{\alpha_m}) = f_0'(x_{\alpha_1}) f_1'(x_{\alpha_1}, x_{\alpha_2}) \cdots f_{m-1}'(x_{\alpha_{m-1}}, x_{\alpha_m}) G'(x_{\alpha_m}).$$

(Con't) No need to prove Proposition 2.10 for assignment 1. This Proof is included for your self study only.

Then apply Proposition 2.9 to see that $X_{\alpha_1} \to X_{\alpha_2} \to \cdots \to X_{\alpha_m}$ forms a Markov chain.

Proof of Propositions 2.19, 2.21, and 2.22

Consider

$$
\begin{aligned}
H(X) - H(X|Y) &= -E \log p(X) + E \log p(Y|X) \\
&= E \log \frac{p(Y|X)}{p(X)} \\
&= E \log \frac{p(X,Y)}{p(X)p(Y)} \\
&= I(X;Y)
\end{aligned}
$$

This proves the first part of Proposition 2.19. The rest of the proposition as well as Propositions 2.21 and 2.22 can be proved likewise.

3. Give an example which shows that pairwise independence does not imply mutual independence.

Solution:

xyz	000	001	010	011	100	101	110	111
p(x,y,z)	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	0

For this joint distribution, X, Y, and Z, are pairwise independent but not mutually independent. Alternatively, the joint distribution for X, Y, and Z can be described by $Z = X + Y \bmod 2$, where X and Y are independent and identical with uniform distribution on $\{0, 1\}$.

4. Verify that $p(x,y,z)$ as defined in Definition 2.4 is a probability distribution. You should exclude all the zero probability masses from the summation carefully.

Solution:

Consider

$$\begin{aligned}
\sum_{x,y,z} p(x,y,z) &= \sum_{y \in \mathcal{S}_y} \sum_{x,z} \frac{p(x,y)p(y,z)}{p(y)} \\
&= \sum_{y \in \mathcal{S}_y} \sum_{x,z} p(x,y)p(z|y) \\
&= \sum_{y \in \mathcal{S}_y} \sum_{x} p(x,y) \sum_{z} p(z|y) \\
&= \sum_{y \in \mathcal{S}_y} \sum_{x} p(x,y) \\
&= \sum_{y \in \mathcal{S}_y} p(y) \\
&= 1.
\end{aligned}$$

5. *Linearity of expectation* It is well-known that expectation is linear, i.e., $E[f(X)+g(Y)] = Ef(X)+Eg(Y)$, where the summation in an expectation is taken over the corresponding alphabet. However, we adopt in information theory the convention that the summation in an expectation is taken over the corresponding support. Justify carefully the linearity of expectation under this convention.

   Solution:

   Consider

$$
\begin{aligned}
& E[f(X)+g(Y)] \\
& \quad = \sum_{(x,y)\in\mathcal{S}_{XY}} p(x,y)(f(X)+g(Y)) \\
& \quad = \sum_{(x,y)\in\mathcal{S}_{XY}} p(x,y)f(x) + \sum_{(x,y)\in\mathcal{S}_{XY}} p(x,y)g(y) \\
& \quad = \sum_{x\in\mathcal{S}_x} \sum_{y:(x,y)\in\mathcal{S}_{XY}} p(x,y)f(x) + \sum_{y\in\mathcal{S}_Y} \sum_{x:(x,y)\in\mathcal{S}_{XY}} p(x,y)g(y) \\
& \quad = \sum_{x\in\mathcal{S}_X} p(x)f(x) + \sum_{y\in\mathcal{S}_Y} p(y)g(y) \\
& \quad = Ef(X)+Eg(Y).
\end{aligned}
$$

   Thus the linearity of the "information-theoretic" expectation operator is justified no matter what values $f(x)$ and $g(y)$ may take (possibly $+\infty$ or $-\infty$) for $x \notin \mathcal{S}_X$ and $y \notin \mathcal{S}_Y$, respectively.

8. Let p_k and p be probability distributions defined on a common finite alphabet. Show that as $k \to \infty$, if $p_k \to p$ in variational distance, then $p_k \to p$ in $\mathcal{L}^2$, and vice versa.

   Solution:

   Note that the variational distance is exactly the $\mathcal{L}^1$-norm. Thus it suffices to show that in $\Re^{|\mathcal{X}|}$, where $|\mathcal{X}|$ is finite, $\mathcal{L}^1$ convergence is equivalent to $\mathcal{L}^2$ convergence. Toward this end, consider any $u = (u(x), x \in \mathcal{X}) \in \Re^{|\mathcal{X}|}$. Then for all $\epsilon > 0$,

$$
\begin{array}{rcl}
 & \sqrt{\sum_x u(x)^2} < \epsilon & \\
\Rightarrow & \sum_x u(x)^2 < \epsilon^2 & \\
\Rightarrow & u(x)^2 < \epsilon & \forall x \in \mathcal{X} \\
\Rightarrow & |u(x)| < \sqrt{\epsilon} & \forall x \in \mathcal{X} \\
\Rightarrow & \sum_x |u(x)| < |\mathcal{X}|\sqrt{\epsilon}. &
\end{array}
$$

   Thus we have shown that $u \to 0$ (the zero vector) in $\mathcal{L}^2$ implies $u \to 0$ in $\mathcal{L}^1$. Similarly, it can be shown that $u \to 0$ in $\mathcal{L}^1$ implies $u \to 0$ in $\mathcal{L}^2$. The proof is completed upon letting $u(x) = p_k(x) - p(x)$ for $x \in \mathcal{X}$ and $k \to \infty$.