Final Exam - Moed B (Solutions)
Total time for the exam: 3 hours!

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1. **True or False (20 Points).**

(a) If $X-Y-Z-W$ is a Markov chain, then $X-Y-Z$ and $Y-Z-W$ are Markov chains.
Solution: **True.**
Since $X-Y-Z-W$ is a Markov chain, then we readily notice that:
1) $P(z|x,y)=P(z|y)$, thus $X-Y-Z$ is a Markov chain.
2) $P(w|y,z)=P(w|z)$, thus $Y-Z-W$ is a Markov chain.

(b) If $X-Y-Z$ and $Y-Z-W$ are Markov chains, then $X-Y-Z-W$ is a Markov chain.
Solution: **False.**
Let (X,Y,Z) be mutually independent, and set $W=X$. Then, obviously, $X-Y-Z$ and $Y-Z-W$ hold, but $X-Y-Z-W$, which reduces to $X-Y-Z-X$, is wrong.

(c) If $P_{XYZW}(x,y,z,w)=P_X(x)P_{Y|X}(y|x)P_{ZW}(z,w)$, then $X-Y-(Z,W)$ is a Markov chain.
Solution: **True.**
Due to the fact that $P_{XYZW}(x,y,z,w)=P_{XY}(x,y)P_{ZW}(z,w)$, we may conclude that $(X,Y)\perp\!\!\!\perp(Z,W)$. Thus, $I(X,Y;Z,W)=0$, from which we can deduce that $I(Y;Z,W|X)=0$, as required.

(d) If $X\perp\!\!\!\perp Y$ and $Y\perp\!\!\!\perp Z$, then $X\perp\!\!\!\perp Z$.
Solution: **False.**
Take $Z=X$.

(e) If the conditional distribution $P_{X|YZ}(x|y,z)$ is a deterministic function of $(x,y)\in\mathcal{X}\times\mathcal{Y}$, then $X-Y-Z$ is a Markov chain.
Solution: **True.**
Since $P_{X|YZ}(x|y,z)$ is a deterministic function of $(x,y)\in\mathcal{X}\times\mathcal{Y}$, then *for any* $z_0\in\mathcal{Z}$ we have

$$P_{X|Y}(x|y)=\sum_{z\in\mathcal{Z}}P_{Z|Y}(z|y)P_{X|YZ}(x|y,z) \tag{1}$$
$$=P_{X|YZ}(x|y,z_0)\sum_{z\in\mathcal{Z}}P_{Z|Y}(z|y) \tag{2}$$
$$=P_{X|YZ}(x|y,z_0). \tag{3}$$

Therefore $X-Y-Z$ is a Markov chain.

2. **Markov with random index (10 points)** Let $A_1,A_2\in\mathcal{A}$, $B_1,B_2\in\mathcal{B}$, and $C\in\mathcal{C}$, be such that A_1-B_1-C and A_2-B_2-C are Markov chains. Also, let $i(C)$ be a binary (deterministic) function of C that emits 1 or 2.

(a) $A_{i(C)}-B_{i(C)}-C$ holds if $P_{A_1,B_1}(a,b)=P_{A_2,B_2}(a,b)$ for all $(a,b)\in\mathcal{A}\times\mathcal{B}$.
Solution: **True.**
Since $P_{A_1,B_1}(a,b)=P_{A_2,B_2}(a,b)$ for every $(a,b)\in\mathcal{A}\times\mathcal{B}$,

$$\begin{aligned}P_{A_1}(a)=&P_{A_2}(a)\\ P_{B_1}(b)=&P_{B_2}(b)\\ P_{A_1|B_1}(a|b)=&P_{A_2|B_2}(a|b)=f(a,b).\end{aligned} \tag{4}$$

Consider

$$\begin{aligned}P_{A_{i(C)},B_{i(C)},C}(a,b,c)=&P_{A_{i(C)}|B_{i(C)},C}(a|b,c)P_{B_{i(C)},C}(b,c)\\ \stackrel{(a)}{=}&P_{A_{i(c)}|B_{i(c)},C}(a|b,c)P_{B_{i(C)},C}(b,c)\\ \stackrel{(b)}{=}&P_{A_{i(c)}|B_{i(c)}}(a|b)P_{B_{i(C)},C}(b,c)\\ \stackrel{(c)}{=}&f(a,b)P_{B_{i(C)},C}(b,c)\end{aligned}$$

where:
(a) - follows since $C=c$ is given,
(b) - follows since $A_{i(c)} - B_{i(c)} - C$ is Markov for every $c \in \mathcal{C}$,
(c) - since $P_{A_{i(c)}|B_{i(c)}}(a|b) = f(a,b)$ for every $c \in \mathcal{C}$.

Therefore, $P_{A_{i(C)}|B_{i(C)},C}(a|b,c) = f(a,b)$, and by Question 1(e) , $A_{i(C)} - B_{i(C)} - C$ is a Markov chain.

(b) $A_{i(C)} - B_{i(C)} - C$ holds if $P_{A_1,B_1}(a,b) \neq P_{A_2,B_2}(a,b)$ for some $(a,b) \in \mathcal{A} \times \mathcal{B}$.
Solution: **False.**
Consider the following example, where $A_1 = 1$, $A_2 = 0$, $B_1 = B_2 = 0$, with probability 1, and let $C \sim \text{Ber}(\alpha)$, for $0 \leq \alpha \leq 1$. Also, assume that (A_1, B_1, C) and (A_2, B_2, C) are mutually independent, and thus

$$\begin{aligned} &A_1 - B_1 - C, \\ &A_2 - B_2 - C, \end{aligned}$$

are Markov chains. Now, let $i(C) \triangleq C + 1$, and define,

$$\begin{aligned} A_{i(C)} &\triangleq \begin{cases} A_1, & \text{if } \ i(C) = 1 \\ A_2, & \text{if } \ i(C) = 2 \end{cases} \\ &= \begin{cases} 1, & \text{if } \ C = 0 \\ 0, & \text{if } \ C = 1 \end{cases}, \end{aligned}$$

and

$$\begin{aligned} B_{i(C)} &\triangleq \begin{cases} B_1, & \text{if } \ i(C) = 1 \\ B_2, & \text{if } \ i(C) = 2 \end{cases} \\ &= 0. \end{aligned}$$

It is clear that $A_{i(C)} - B_{i(C)} - C$ is not Markov, since $B_{i(C)}$ is deterministic and $A_{i(C)}$ depends on C.

3. **Erasure channel after discrete memoryless channel (20 Points):**

Assume a discrete memoryless channel, $(\mathcal{X}, \mathcal{Y}, p(y|x))$ with capacity, C_1.
The output of this channel serves as an input to an *erasure channel* with $|\mathcal{Y}|$ inputs and erasure probability ϵ.
What is the capacity of the overall channel?
Solution: We will show that the capacity, C, of the overall channel is $C = (1-\epsilon) \cdot C_1$. Indeed, since the overall channel is *memoryless*, we saw in the lectures that its capacity is given by

$$C = \max_{P_x} I(X;Z). \tag{5}$$

Define,

$$\Theta \triangleq \begin{cases} 1 & \text{if } z \in \mathcal{Y} \\ 0 & \text{if } z = e \end{cases}.$$

Then,

$$\begin{aligned} &\Pr(\Theta = 1) = \Pr(Z = Y) = 1 - \epsilon, \\ &\Pr(\Theta = 0) = \Pr(Z = e) = \epsilon. \end{aligned} \tag{6}$$

Now, due to the fact that Θ is deterministic function of Z, we get

$$\begin{aligned} I(X;Z) &= I(X;Z,\Theta) \\ &= H(X) - H(X|Z,\Theta) \\ &= H(X) - \Pr(\Theta = 1) \cdot H(X|Z,\Theta = 1) - \Pr(\Theta = 0) \cdot H(X|Z,\Theta = 0) \\ &= H(X) - (1-\epsilon) \cdot H(X|Y) - \epsilon \cdot H(X|Z = e) \\ &= H(X) - (1-\epsilon) \cdot H(X|Y) - \epsilon \cdot H(X) \\ &= (1-\epsilon) \cdot I(X;Y). \end{aligned} \tag{7}$$

Therefore, the capacity of the overall channel is:

$$\begin{aligned} C &= \max_{P_x} I(X;Z) \\ &= (1-\epsilon)\cdot \max_{P_x} I(X;Y) \\ &= (1-\epsilon)\cdot C_1, \end{aligned} \tag{8}$$

as claimed.

4. **Cross entropy (25 Points)**:

Often in Machine learning, cross entropy is used to measure performance of a classifier model such as neural network. Cross entropy is defined for two PMFs P_X and Q_X as

$$H(P_X, Q_X) \triangleq -\sum_{x\in\mathcal{X}} P_X(x)\log Q_X(x).$$

In a shorter notation we write as

$$H(P, Q) \triangleq -\sum_{x\in\mathcal{X}} P(x)\log Q(x).$$

(a) Copy each of the following relations to your notebook and write **true** ir **false** and provide a proof/disproof.

i. $0 \le H(P,Q) \le \log|\mathcal{X}|$ for all P, Q.
Solution: **False.**
First, note that $H(P,Q)$ can be rewritten as

$$\begin{aligned} H(P,Q) &= -\sum_{x\in\mathcal{X}} P(x)\log Q(x) \\ &= \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} - \sum_{x\in\mathcal{X}} P(x)\log P(x) \\ &= D(P||Q) + H_P(X). \end{aligned} \tag{9}$$

Thus, it obvious that $H(P,Q) \ge 0$. However, if we let P_{unif} be the uniform measure on $\mathcal{X}$, then

$$\begin{aligned} H(P_{\text{unif}}, Q) &= D(P_{\text{unif}}||Q) + H_{P_{\text{unif}}}(X) \\ &= D(P_{\text{unif}}||Q) + \log|\mathcal{X}| \\ &\ge \log|\mathcal{X}|, \end{aligned} \tag{10}$$

due to the fact that $D(P_{\text{unif}}||Q) \ge 0$. Now, because $D(P_{\text{unif}}||Q) = 0$ if and only if $Q = P_{\text{unif}}$, by taking any $Q \ne P_{\text{unif}}$, we will get that $D(P_{\text{unif}}||Q) > 0$, which means that $H(P_{\text{unif}}, Q) > \log|\mathcal{X}|$ for any $Q \ne P_{\text{unif}}$, contradicting the claim that $H(P,Q) \le \log|\mathcal{X}|$ for all P, Q.

ii. $\min_Q H(P,Q) = H(P,P)$ for all P.
Solution: **True.**
This follows from the simple observation that $D(P||Q) \ge 0$ for all (P,Q), and thus

$$\begin{aligned} H(P,Q) &= D(P||Q) + H_P(X) \\ &\ge H_P(X), \end{aligned} \tag{11}$$

with equality if and only if $Q = P$.

iii. $H(P,Q)$ is concave in the pair (P,Q).
Solution: **False.**
If $H(P,Q)$ is concave in the pair (P,Q) then it must be concave in P and Q separately. However, it easy to see that $H(P,Q)$ is convex function in Q (for fixed P) because $-\log(\cdot)$ is convex.

iv. $H(P,Q)$ is convex in the pair (P,Q).
Solution: **False.**
If $P = Q$, then $H(P,Q) = H_P(X)$, which is a concave function of P.

(b) Find an operation problem, such as in compression, communication (or even other fields) where the fundamental solution involve the cross entropy measure $H(P,Q)$. State the operational problem mathematically in less than half a page, and state the solution as a theorem. Provide a short proof to the theorem.
Solution: Cross entropy is the expected length of a code, when using the distribution Q to encode a source with distribution P. For example, consider the Shannon code length $L(x) = \lceil \log 1/Q(x) \rceil$. However, consider the case

where the true p.m.f. is P. Therefore, one should expect that we will not achieve an expected (optimal) length of $H(P)$. Instead, we will get

$$\begin{aligned}\mathbb{E}(L(X)) &= \sum_x P(x)\left\lceil \log\frac{1}{Q(x)}\right\rceil \\ &\leq 1 + \sum_x P(x)\log\frac{1}{Q(x)} \\ &= 1 + H(P,Q),\end{aligned} \tag{12}$$

and similarly

$$\mathbb{E}(L(X)) \geq H(P,Q). \tag{13}$$

Hence,

$$H(P,Q) \leq \mathbb{E}(L(X)) \leq 1 + H(P,Q). \tag{14}$$

5. **Fast fading Gaussian channel (25 points)**:
Consider a Gaussian channel given by $Y_i = G_i X_i + Z_i$, where $Z_i \overset{i.i.d}{\sim} \mathcal{N}(0,N)$ and $G_i \overset{i.i.d}{\sim} P_G$.

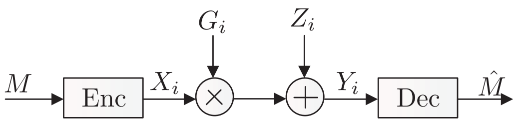

Figure 1: Fast fading Gaussian channel

The gains and noise are independent, i.e., $\{Z_i\} \perp\!\!\!\perp \{G_i\}$, and

$$P_G(g) = \begin{cases} 0.5 & \text{if } g = 1 \\ 0.5 & \text{if } g = 2 \end{cases}$$

(a) Assume that the states are known at the decoder only, and there is an input constraint P.

i. What is the capacity formula?
Solution: As we saw in the lectures, the capacity is given by

$$C_1 = \sup_{P_X} I(X;Y|G) \tag{15}$$

where the maximum is taken over all X-distributions such that $E(X^2) \leq P$.

ii. Find the optimal inputs distribution in the formula you gave.
Solution: We show in the next item that the optimal input is $X \sim \mathcal{N}(0,P)$.

iii. Compute the capacity as a function of N and P.
Solution: The states are known only to the decoder, and thus we may assume that $X \perp\!\!\!\perp G$. We have

$$\begin{aligned} I(X;Y|G) &= h(Y|G) - h(Y|X,G) \\ &= h(Y|G) - h(Y - GX|X,G) \\ &= h(Y|G) - h(Z|X,G) \\ &\overset{(a)}{=} h(Y|G) - h(Z) \\ &= P(G=1)\cdot h(Y|G=1) + P(G=2)\cdot h(Y|G=2) - h(Z) \\ &= \frac{1}{2}\cdot h(X+Z|G=1) + \frac{1}{2}\cdot h(2X+Z|G=2) - h(Z) \\ &\overset{(a)}{=} \frac{1}{2}\cdot h(X+Z) + \frac{1}{2}\cdot h(2X+Z) - \frac{1}{2}\log(2\pi e N) \end{aligned} \tag{16}$$

where (a) follows from the fact that $Z \perp\!\!\!\perp (X,G)$. Now, the maximal entropy lemma implies

$$h(X+Z) \leq \frac{1}{2}\log(2\pi e(P+N)), \tag{17}$$

$$h(2X+Z) \leq \frac{1}{2}\log(2\pi e(4P+N)), \tag{18}$$

with *equality* if and only if $X \sim \mathcal{N}(0, P)$. Thus,

$$I(X;Y|G) \leq \frac{1}{2}\log\left(\frac{\sqrt{(P+N)(4P+N)}}{N}\right), \tag{19}$$

again, with *equality* if and only if $X \sim \mathcal{N}(0, P)$. Therefore

$$\begin{aligned} C_1 &= \max_{P_X} I(X;Y|G) \\ &= \frac{1}{4}\log\left(1+\frac{P}{N}\right)+\frac{1}{4}\log\left(1+\frac{4P}{N}\right). \end{aligned} \tag{20}$$

(b) i. Now the states are known both to the encoder and decoder, and the input constraint is P.

A. What is the capacity formula?

Solution: Again, as we saw in the lectures, the capacity in this case is given by

$$C_2 = \sup_{P_{X|G}} I(X;Y|G) \tag{21}$$

where the maximum is over all distributions $P_{X|G}$ which satisfy the power constraint.

B. Compute the capacity as a function of N and P.
You can write your answer as an optimization problem.

Solution: As before, we get:

$$I(X;Y|G) = \frac{1}{2}\cdot h(X+Z|G=1)+\frac{1}{2}\cdot h(2X+Z|G=2)-\frac{1}{2}\log(2\pi eN). \tag{22}$$

Define the functional:

$$\text{var}(X|W=w) \triangleq \mathbb{E}(X^2|W=w). \tag{23}$$

Then, let $\text{var}(X|G=i) \triangleq P_i$ for $i = 1, 2$. By the power constraint, we have that $P_1 + P_2 \leq 2P$. Accordingly, due to the maximal entropy lemma, we get

$$h(X+Z|G=1) \leq \frac{1}{2}\log(2\pi e(P_1+N)) \tag{24}$$

$$h(2X+Z|G=2) \leq \frac{1}{2}\log(2\pi e(4P_2+N)), \tag{25}$$

were both inequalities are achieved if $X|G=i$ is Gaussian with variance P_i. Hence,

$$I(X;Y|G) \leq \frac{1}{2}\log\left(\frac{\sqrt{(P_1+N)(P_2+N)}}{N}\right). \tag{26}$$

Therefore, the capacity is given by

$$\begin{aligned} C_2 &= \sup_{(P_1,P_2):P_1+P_2\leq 2P} \frac{1}{2}\log\left(\frac{\sqrt{(P_1+N)(4P_2+N)}}{N}\right) \qquad (27) \\ &= \sup_{(P_1,P_2):P_1+P_2\leq 2P} \left\{\frac{1}{4}\log\left(1+\frac{P_1}{N}\right)+\frac{1}{4}\log\left(1+\frac{4P_2}{N}\right)\right\}. \qquad (28) \end{aligned}$$

Remark: This optimization problem can be solved by Water-filling.

ii. Assume

$$P_G(g) = \begin{cases} 0.5 & \text{if } g=0 \\ 0.5 & \text{if } g=1 \end{cases}.$$

Repeat 5(b)i.

Solution: As before we will get

$$\begin{aligned} I(X;Y|G) &= \frac{1}{2}\cdot h(Z|G=0)+\frac{1}{2}\cdot h(X+Z|G=1)-\frac{1}{2}\log(2\pi eN) \\ &= \frac{1}{2}\cdot h(X+Z|G=1)-\frac{1}{4}\log(2\pi eN). \end{aligned} \tag{29}$$

Also,

$$h(X+Z|G=1) \leq \frac{1}{2}\log(2\pi e(P_1+N)). \tag{30}$$

Thus,

$$\begin{aligned} C_3 &= \sup_{(P_1):P_1\leq 2P} \frac{1}{4}\log\left(\frac{P_1+N}{N}\right) \\ &= \frac{1}{4}\log\left(1+\frac{2P}{N}\right). \end{aligned} \tag{31}$$

Intuition: To achieve (31) the transmitter will not send any data when $G=0$ (because in this case our information will be lost), but rather all the data will be transmitted when $G=1$ (and with power $2P$ to satisfy the power constraint). Now, when $G=1$, the channel reduces to a simple Gaussian channel, $Y=X+Z$, with signal to noise ratio of $2P/N$, namely, we can achieve a rate of

$$\frac{1}{2}\log\left(1+\frac{2P}{N}\right).$$

However, since with high probability, $G=1$ half of the time, we need to multiply the last result by half, and we get the quarter factor in (31).

Good Luck!