EE376A: Final Solutions

1. Channel Capacity in Presence of State Information (*32 points*)
Consider a binary memoryless channel with state whose input-output relation is described as follows. The received channel output Y is given by

$$Y = X \oplus S \oplus Z,$$

where X is the channel input, S is the channel state, and Z is the channel noise. X, S, Z and Y all take their values in $\{0, 1\}$, and $\oplus$ denotes modulo-2 addition. $S \sim \text{Bernoulli}(q)$ and $Z \sim \text{Bernoulli}(p)$ are independent, and jointly independent of the channel input X. Thus, when we employ n-block encoding and decoding, we have for each $1 \leq i \leq n$

$$Y_i = X_i \oplus S_i \oplus Z_i,$$

where S^n are i.i.d. Bernoulli(q) and Z^n are i.i.d. Bernoulli(p), where S^n and Z^n are independent, and jointly independent of the channel input sequence X^n.

Find the capacity of the channel in the following cases:

(a) (*8 points*) Neither the transmitter nor the receiver knows the state sequence.

(b) (*8 points*) The state sequence is known only to the receiver, i.e., the n-block decoder gets to base its decision on Y^n and S^n.

(c) (*8 points*) The state sequence is known to both the transmitter and the receiver, i.e., the n-block encoder and decoder know S^n prior to communication.

(d) (*8 points*) The state sequence is known to both the transmitter and the receiver, as in the previous part but, in addition, the encoding must adhere to the cost constraint $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] \leq 0.25$.

Solution:

(a) The state can be considered as a part of noise, i.e., $S \oplus Z$ can be treated as a noise. Therefore, the channel is a binary symmetric channel with crossover probability $p * q = p(1-q) + (1-p)q$, ant the capacity is $C_a = 1 - h_2(p * q)$ where $h_2(x)$ is a binary entropy function.

(b) The state can be considered as a part of output, i.e., (S, Z) can be treated as an output. Therefore, the capacity is

$$\begin{aligned} C_b &= \max_{P_X} I(X; Y, S) \\ &= \max_{P_X} H(Y, S) - H(Y, S|X) \\ &= \max_{P_X} H(X \oplus Z, S) - H(Z, S) \\ &= \max_{P_X} H(X \oplus Z) + H(S) - H(Z) - H(S) \\ &= 1 - h_2(p). \end{aligned}$$

where the capacity achieving distribution is Bernoulli($\frac{1}{2}$).

(c) Since the state is given to both the encoder and the decoder, the capacity is

$$\begin{aligned} C_c &= \max_{P_{X|S}} I(X;Y|S) \\ &= \max_{P_{X|S}} H(Y|S) - H(Y|X,S) \\ &= \max_{P_{X|S}} H(X \oplus Z|S) - H(Z) \\ &= 1 - h_2(p). \end{aligned}$$

(d) The capacity with cost constraint is

$$\begin{aligned} C_d &= \max_{P_{X|S}:\mathbb{E}[X]\leq 0.25} I(X;Y|S) \\ &= \max_{P_{X|S}:\mathbb{E}[X]\leq 0.25} H(Y|S) - H(Y|X,S) \\ &= \max_{P_{X|S}:\mathbb{E}[X]\leq 0.25} H(X \oplus Z|S) - H(Z) \\ &\leq \max_{P_{X|S}:\mathbb{E}[X]\leq 0.25} H(X \oplus Z) - H(Z) \\ &= \max_{P_X:\mathbb{E}[X]\leq 0.25} H(X \oplus Z) - H(Z) \\ &= h_2(\frac{1}{4} * p) - h_2(p). \end{aligned}$$

It is clear that the equality can be achieved by $X \sim Bern(\frac{1}{4})$ that are independent to S. Therefore,

$$C_d = h_2(\frac{1}{4} * p) - h_2(p).$$

2. Rate Distortion under Log Loss (*40 points*)

Let X be distributed i.i.d. according to distribution $P \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability mass functions on the finite alphabet $\mathcal{X}$. Distortion functions usually ask for "hard" reconstructions. For example, the reconstruction alphabet is the same as the source alphabet, $\mathcal{Y} = \mathcal{X}$, and the distortion is Hamming $d(x,y) = \mathbf{1}(x \neq y)$.

Suppose we are instead interested in a "soft" reconstruction: instead of a finite reconstruction alphabet we let $\mathcal{Y} = \mathcal{P}(\mathcal{X})$, i.e., the reconstruction is a PMF on $\mathcal{X}$. Thus, a reconstruction $q \in \mathcal{P}(\mathcal{X})$ may be construed as a distribution of belief $q(x)$ by the decoder over the values $x \in \mathcal{X}$ that the source symbol might take. A natural distortion function for such soft reconstructions is the 'log-loss':

$$d_\ell(x,q) = -\log q(x).$$

(a) (*8 points*) Show that

$$\mathbb{E}[d_\ell(X, q)] \geq H(X),$$

for all $q \in \mathcal{P}(\mathcal{X})$. When is equality achieved?

(b) (*8 points*) Suppose X and U are jointly distributed and denote the conditional PMF of X given U by $P_{X|U}$. Show that

$$\mathbb{E}\left[d_\ell\left(X, P_{X|U}(\cdot|U)\right)\right] = H(X|U),$$

where $P_{X|U}(\cdot|u)$ denotes the conditional PMF of X given $U = u$ and, therefore, $P_{X|U}(\cdot|U)$ is a $\mathcal{P}(\mathcal{X})$-valued random variable.

(c) (*8 points*) Consider the random PMF Q — that is, it is a random variable on $\mathcal{P}(\mathcal{X})$. Let Q be jointly distributed with X according to $P_{X,Q}$. Furthermore, let $\tilde{Q} = P_{X|Q}(\cdot|Q)$ where $\tilde{Q}$ is again a random variable on $\mathcal{P}(\mathcal{X})$. Show that

$$\mathbb{E}\left[d_\ell(X, Q)\right] \geq \mathbb{E}\left[d_\ell(X, \tilde{Q})\right] = H(X|Q).$$

Also, argue why

$$I(X; \tilde{Q}) \leq I(X; Q).$$

(d) (*8 points*) Find and plot the rate distortion function under log-loss. I.e., find

$$R(D) = \min_{\mathbb{E}[d_\ell(X,Q)] \leq D} I(X; Q).$$

Hint: use the previous part to show that the constraint set for the minimization can be taken to be $H(X|Q) \leq D$ instead of $\mathbb{E}[d_\ell(X, Q)] \leq D$.

(e) (*8 points*) For a given distortion level D, describe a concrete implementable scheme (not based on a random coding argument) for achieving $R(D)$.

Solution: Rate Distortion under Log Loss.

(a) Since X is distributed according to P,

$$\begin{aligned}
\mathbb{E}[d_\ell(X, q)] - H(X) &= \mathbb{E}[-\log q(X)] - \mathbb{E}[-\log P(X)] \\
&= \mathbb{E}[\log \frac{P(X)}{q(X)}] \\
&= D(P||q) \\
&\geq 0.
\end{aligned}$$

The equality holds if and only if $q = P$.

(b) By definition of $d_\ell(\cdot,\cdot)$ and the tower property,

$$\begin{aligned}\mathbb{E}[d_\ell(X, P_{X|U}(\cdot|U))] &= \mathbb{E}[\mathbb{E}[d_\ell(X, P_{X|U}(\cdot|U))|U]] \\ &= \mathbb{E}[\mathbb{E}[-\log P_{X|U}(X|U)|U]] \\ &= H(X|U).\end{aligned}$$

(c) By the tower property,

$$\begin{aligned}\mathbb{E}[d_\ell(X, Q)] &= \mathbb{E}[\mathbb{E}[d_\ell(X, Q)|Q]] \\ &\geq \mathbb{E}[\mathbb{E}[d_\ell(X, P_{X|Q}(\cdot|Q))|Q]] \\ &= \mathbb{E}[d_\ell(X, P_{X|Q}(\cdot|Q))] \\ &= \mathbb{E}[d_\ell(X, \tilde{Q})] \\ &= H(X|Q).\end{aligned}$$

Note that $\mathbb{E}[d_\ell(X, Q)] \geq H(X)$ is not true for random pmf Q.
Since $\tilde{Q}$ is a function of Q,

$$H(X|Q) \leq H(X|\tilde{Q}).$$

Therefore,

$$I(X;\tilde{Q}) = H(X) - H(X|\tilde{Q}) \leq H(X) - H(X|Q) \leq I(X;Q).$$

(d) In part (c), we have seen that $\mathbb{E}[d_\ell(X, Q)] \leq D$ implies $H(X|Q) \leq D$. Therefore,

$$\begin{aligned}R(D) &= \min_{\mathbb{E}[d_\ell(X,Q)]\leq D} I(X;Q) \\ &\geq \min_{H(X|Q)\leq D} I(X;Q) \\ &= \min_{H(X|Q)\leq D} H(X) - H(X|Q) \\ &= H(X) - D.\end{aligned}$$

On the other hand, since $\mathbb{E}[d_\ell(X, \tilde{Q})] = H(X|Q)$, we have

$$\begin{aligned}H(X) - D &= \min_{H(X|Q)\leq D} I(X;Q) \\ &\geq \min_{\mathbb{E}[d_\ell(X,\tilde{Q})]\leq D} I(X;\tilde{Q}) \\ &= R(D).\end{aligned}$$

Therefore, $R(D) = H(X) - D$ which is a straight line between $(H(X), 0)$ and $(0, H(X))$.

(e) Since the rate-distortion curve is a straight line, we can use the time sharing argument. First, we have a concrete scheme that achieves that losslessly compress the source using the rate $H(X)$ (Enumerating all typical sequences). Also, we have a concrete scheme that achieves the distortion $H(X)$ using zero-rate (the reconstruction is simply PMF of X).
By using first scheme for $\frac{H(X)-D}{H(X)}$ fraction of time and using second scheme for $\frac{D}{H(X)}$ fraction of time, we can achieve the distortion D with a rate $R = H(X) - D$.

3. The MMI decoder (*40 points*)

Valery has discovered an amazing new channel decoder. He claims it needs to know nothing about the channel! Imre and Janos are suspicious and need your help checking the validity of his claim.

The (standard random coding) assumptions:

- 2^{nR} codewords are drawn i.i.d. according to distribution P_X.
- Message J is chosen uniformly from the set $\{1, 2, \ldots, 2^{nR}\}$, and the corresponding codeword $X^n(J)$ is sent through the channel.
- The channel is discrete memoryless, characterized by the conditional PMF $P_{Y|X}$, where both X and Y take values in the respective finite alphabets $\mathcal{X}$ and $\mathcal{Y}$.

Now suppose sequence $Y^n \in \mathcal{Y}^n$ is received by the decoder. Let $P_{X^n(j),Y^n}$ denote the joint empirical distribution of the jth codeword $X^n(j)$ and Y^n. Valery's decoder produces as its estimate the codeword with maximum empirical mutual information with Y^n:

$$\hat{J} = \underset{j \in \{1,2,\ldots,2^{nR}\}}{\arg\max} \; I(P_{X^n(j),Y^n})$$

where, for $Q_{X,Y} \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$, $I(Q_{X,Y})$ denotes the mutual information between X and Y when distributed according to $Q_{X,Y}$ (and ties in the maximization are broken arbitrarily).

(a) (*8 points*) Does Valery's decoder have to know the channel statistics $P_{Y|X}$ in order to implement this decoding scheme?

(b) (*8 points*) Prove that, for any $Q_{X,Y} \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$,

$$D\left(Q_{X,Y} \| P_X \times P_Y\right) \geq I(Q_{X,Y}).$$

Hint: Use the fact that $I(Q_{X,Y}) = D\left(Q_{X,Y} || Q_X \times Q_Y\right)$.

(c) (*8 points*) Using the previous part, prove that $f(\theta) \geq \theta$, where

$$f(\theta) \stackrel{\Delta}{=} \min_{Q_{X,Y} \in \mathcal{P}(\mathcal{X},\mathcal{Y}): I(Q_{X,Y}) \geq \theta} D\left(Q_{X,Y} \| P_X \times P_Y\right).$$

(d) (*8 points*) Using the method of types show that for any $2 \leq j \leq 2^{nR}$

$$\mathsf{P}\left(I(P_{X^n(j),Y^n}) \geq \theta \middle| \, J = 1\right) \doteq 2^{-nf(\theta)}$$

(e) (*8 points*) What is the supremum of rates R for which

$$\mathsf{P}\left(\hat{J} \neq J\right) \longrightarrow 0 \quad \text{as} \quad n \to \infty$$

under Valery's scheme? How does it compare to the supremum of rates for which joint typicality decoding would have achieved reliable communication?

Hint: First justify why for any θ

$$\begin{aligned}
& \mathsf{P}\left(\hat{J} \neq J\right) \\
= \; & \mathsf{P}\left(\hat{J} \neq J | J=1\right) \\
\leq \; & \mathsf{P}\left(I(P_{X^n(1),Y^n}) < \theta \middle| \, J=1\right) + \mathsf{P}\left(I(P_{X^n(j),Y^n}) \geq \theta \text{ for some } 2 \leq j \leq 2^{nR} \middle| \, J=1\right) \\
\leq \; & \mathsf{P}\left(I(P_{X^n(1),Y^n}) < \theta \middle| \, J=1\right) + 2^{nR} \times \mathsf{P}\left(I(P_{X^n(2),Y^n}) \geq \theta \middle| \, J=1\right).
\end{aligned}$$

Then, with the help of previous parts, establish for which values of R and θ the expression in the last line vanishes.

Solution:

(a) It only needs to know the $X^n(j)$ for all $1 \leq j \leq 2^{nR}$ and the output Y^n. Unlike to the joint typicality decoding, the decoder does not have to know about the channel statistics $P_{Y|X}$.

(b)

$$\begin{aligned}
& D(Q_{X,Y}||P_X \times P_Y) - I(Q_{X,Y}) \\
= \; & \sum_{x,y} Q_{X,Y}(x,y) \log \frac{Q_{X,Y}(x,y)}{P_X(x)P_Y(y)} - \sum_{x,y} Q_{X,Y}(x,y) \log \frac{Q_{X,Y}(x,y)}{Q_X(x)Q_Y(y)} \\
= \; & \sum_{x,y} Q_{X,Y}(x,y) \log \frac{Q_X(x)Q_Y(y)}{P_X(x)P_Y(y)} \\
= \; & \sum_{x} Q_X(x) \log \frac{Q_X(x)}{P_X(x)} + \sum_{y} Q_Y(y) \log \frac{Q_Y(y)}{P_Y(y)} \\
= \; & D(Q_X||P_X) + D(Q_Y||P_Y) \\
\geq \; & 0.
\end{aligned}$$

(c)

$$\begin{aligned}
f(\theta) \; & \triangleq \; \min_{Q_{X,Y} \in \mathcal{P}(\mathcal{X},\mathcal{Y}): I(Q_{X,Y}) \geq \theta} D\left(Q_{X,Y} \| P_X \times P_Y\right) \\
& \geq \; \min_{Q_{X,Y} \in \mathcal{P}(\mathcal{X},\mathcal{Y}): I(Q_{X,Y}) \geq \theta} I\left(Q_{X,Y}\right) \\
& \geq \; \theta.
\end{aligned}$$

(d) Let $E_n = \{Q_{X,Y} \in \mathcal{P}(\mathcal{X},\mathcal{Y}) : I(Q_{X,Y}) \geq \theta\}$. For $j \geq 2$, the joint law of $X^n(j)$ and

Y^n is $P_X \times P_Y$. Therefore,

$$\begin{aligned}
\mathsf{P}\left(I(P_{X^n(j),Y^n}) \geq \theta \middle| J=1\right) &= \mathsf{P}\left(P_{X^n(j),Y^n} \in E_n | J=1\right) \\
&= \sum_{Q_{X,Y} \in E_n} \mathsf{P}(T(Q_{X,Y})) \\
&= \sum_{Q_{X,Y} \in E_n} (P_X \times P_Y)^n (T(Q_{X,Y})) \\
&\doteq 2^{-n \min_{Q_{X,Y} \in E_n} D(Q_{X,Y} || P_X \times P_Y)} \\
&\doteq 2^{-nf(\theta)}
\end{aligned}$$

(e) Recall the hint:

$$\begin{aligned}
& \mathsf{P}\left(\hat{J} \neq J\right) \\
\overset{(i)}{=} \; & \mathsf{P}\left(\hat{J} \neq J | J=1\right) \\
\overset{(ii)}{\leq} \; & \mathsf{P}\left(I(P_{X^n(1),Y^n}) < \theta \middle| J=1\right) + \mathsf{P}\left(I(P_{X^n(j),Y^n}) \geq \theta \text{ for some } 2 \leq j \leq 2^{nR} \middle| J=1\right) \\
\overset{(iii)}{\leq} \; & \mathsf{P}\left(I(P_{X^n(1),Y^n}) < \theta \middle| J=1\right) + 2^{nR} \times \mathsf{P}\left(I(P_{X^n(2),Y^n}) \geq \theta \middle| J=1\right).
\end{aligned}$$

This is because,

- (i) is because of symmetry.
- (ii) is because: $\hat{J} \neq 1$ if $I(P_{X^n(1),Y^n}) < \theta$ or $I(P_{X^n(J),Y^n}) \leq I(P_{X^n(1),Y^n})\theta$ for some other $J \neq 1$.
- (iii) is because of symmetry again.

Suppose that the inequalities $R < \theta < I(X;Y) = I(P_{X,Y})$ hold. Then, by the law of large number,

$$\lim_{n \to \infty} I(P_{X^n(1),Y^n}) = I(P_{X,Y}) > \theta.$$

This implies that $\lim_{n\to\infty} \mathsf{P}\left(I(P_{X^n(1),Y^n}) < \theta \middle| J=1\right) = 0$.
On the other hand, by part (c) and (d),

$$\begin{aligned}
2^{nR} \times \mathsf{P}\left(I(P_{X^n(2),Y^n}) \geq \theta \middle| J=1\right) &\doteq 2^{n(R-f(\theta))} \\
&\leq 2^{n(R-\theta)}
\end{aligned}$$

which vanishes as n grows.

Thus, we can argue that $\mathsf{P}\left(\hat{J} \neq J\right)$ converges to zero as n grows. Finally, using this scheme, we can achieve any rate below $\sup_{P_X} I(X;Y)$ which is the channel capacity that we achieved using joint typicality.