EE326 Introduction to Information Theory and Coding

- TA in charge of Homework #6: Hyeonseo Park (phseo2000@kaist.ac.kr)

9.1

(a) Using the definition of mutual information and the chain rule for entropy we can expand the term $I(X; Y_1, Y_2)$ as follows:

$$\begin{aligned}
I(X; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X) \\
&= H(Y_1) + H(Y_2|Y_1) - (H(Y_1|X) + H(Y_2|Y_1, X)) \\
&= H(Y_1) + H(Y_2) - I(Y_1; Y_2) - (H(Y_1|X) + H(Y_2|X) - I(Y_1; Y_2|X)) \\
&= I(X; Y_1) + I(X; Y_2) - I(Y_1; Y_2) + I(Y_1; Y_2|X) \\
&\overset{(a)}{=} I(X; Y_1) + I(X; Y_2) - I(Y_1; Y_2) \\
&\overset{(b)}{=} 2I(X; Y_1) - I(Y_1; Y_2)
\end{aligned}$$

where (a) follows from the conditional independence of Y_1 and Y_2 given X, which implies $I(Y_1; Y_2|X) = 0$. And (b) follows from the fact that Y_1 and Y_2 are conditionally identically distributed given X, meaning $I(X; Y_1) = I(X; Y_2)$.

(b) The capacity of the single output channel, denoted as C_1, is defined as:

$$C_1 = \max_{p(x)} I(X; Y_1).$$

The capacity of the channel with two independent looks, C_2, can be derived as:

$$\begin{aligned}
C_2 &= \max_{p(x)} I(X; Y_1, Y_2) \\
&= \max_{p(x)} 2I(X; Y_1) - I(Y_1; Y_2) \quad \textbf{(by (a))} \\
&\leq \max_{p(x)} 2I(X; Y_1) \quad \textbf{(mutual information is nonnegative)} \\
&= 2C_1
\end{aligned}$$

In conclusion, the capacity of the channel with two independent looks is less than or equal to twice the capacity of the single output channel.

9.8

We wish to maximize the total capacity subject to the total cost constraint. The optimization problem is formulated as:

$$C = \max_{P_1, P_2, \beta_1 P_1 + \beta_2 P_2 \leq \beta} \frac{1}{2} \log(1 + \frac{P_1}{N_1}) + \frac{1}{2} \log(1 + \frac{P_2}{N_2}).$$

To solve this constrained optimization problem, we use the method of Lagrange multipliers. We define the objective function L as:

$$L(P_1, P_2, \lambda) = \sum_i \frac{1}{2} \ln(1 + \frac{P_i}{N_i}) - \lambda(\sum_i \beta_i P_i - \beta).$$

Now, we differentiate L with respect to P_i and λ and set the derivative to zero to find the optimal power allocation.

$$\frac{\partial L}{\partial P_i} = \frac{1}{2} \cdot \frac{1}{P_i + N_i} - \beta_i \lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = -(\sum_i \beta_i P_i - \beta) = 0$$

Rearranging these equations gives

$$\beta_i P_i + \beta_i N_i = \frac{1}{2\lambda} \quad \text{and} \quad \sum_i \beta_i P_i = \beta.$$

Let $\frac{1}{2\lambda} = \nu$. Then

$$\beta_i P_i + \beta_i N_i = \nu.$$

Since the allocated power must be nonnegative, the optimal power allocation satisfies

$$\beta_i P_i = (\nu - \beta_i N_i)^+ \text{ where } \nu \text{ satisfies } \sum_i (\nu - \beta_i N_i)^+ = \beta.$$

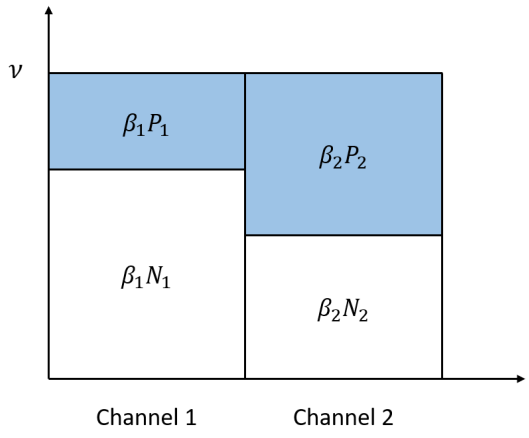

Figure 1: Power allocation of parallel gaussian channels.

(a) Without loss of generality, assume $\beta_1 N_1 \geq \beta_2 N_2$.

- If β is small, all the power is allocated to channel 2.
- We start using channel 1 when ν rises above $\beta_1 N_1$ (see Figure 1).

Therefore, the condition for using both channels is

$$\beta > |\beta_1 N_1 - \beta_2 N_2|.$$

(b) First, check if both channels are used:

$$|\beta_1 N_1 - \beta_2 N_2| = 1 < \beta$$

Thus, both channels will be used.
Next, we solve for ν using $\sum_i (\nu - \beta_i N_i)^+ = \beta$.

$$\begin{aligned} \nu - 3 + \nu - 4 &= 10 \\ \nu &= 8.5 \end{aligned}$$

Now, we calculate P_1 and P_2.

$$\begin{aligned} \beta_1 P_1 = 8.5 - 3 \quad &\Rightarrow \quad P_1 = 5.5 \\ \beta_2 P_2 = 8.5 - 4 \quad &\Rightarrow \quad P_2 = 2.25 \end{aligned}$$

Finally, the channel capacity is

$$C = \frac{1}{2}\log(1 + \frac{5.5}{3}) + \frac{1}{2}\log(1 + \frac{2.25}{2}) \approx 1.295 \text{bits/channel use.}$$

9.3

First, we need to convert the output power constraint into an input power constraint.

$$\begin{aligned} E[Y^2] &= E[(X+Z)^2] \\ &= E[X^2] + 2E[X]E[Z] + E[Z^2] \quad \textbf{(X and Z are independent)} \\ &= E[X^2] + Var[Z] \quad \textbf{(Z has zero mean)} \\ &= E[X^2] + N \\ &\leq P \end{aligned}$$

Thus, the input power constraint is

$$E[X^2] \leq P - N.$$

In conclusion, the channel capacity is

$$\begin{aligned} C &= \frac{1}{2}\log(1 + \frac{P-N}{N}) \\ &= \frac{1}{2}\log(\frac{P}{N}). \end{aligned}$$

10.5

The rate distortion function is

$$R(D) = \min_{p(\hat{x}|x), E[d(x,\hat{x})] \le D} I(X;\hat{X}).$$

We analyze $R(D)$ in two cases based on the range of D.

(i) $D \ge (m-1)/m$

The expected distortion for random guess is $(m-1)/m$. Thus, if the allowed distortion D is greater than or equal to $(m-1)/m$, we can choose $\hat{X}$ to be statistically independent of X (for example, setting $\hat{X} = 1$). In this case, $I(X;\hat{X}) = 0$ and we have $R(D) = 0$.

(ii) $D \le (m-1)/m$

First we derive a lower bound for $I(X;\hat{X})$ using Fano's inequality.

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= \log m - H(X|\hat{X}) \quad \textbf{(}X \textbf{ is uniform)} \\ &\ge \log m - [H_B(P(X \ne \hat{X})) + P(X \ne \hat{X}) \log(m-1)]. \end{aligned}$$

where H_B is the binary entropy function.

Next, we find the maximum of $H_B(P(X \ne \hat{X})) + P(X \ne \hat{X}) \log(m-1)$. Let $q = P(X \ne \hat{X})$ and $f(q) = H_B(q) + q \log(m-1)$. For Hamming distortion, the expected distortion is equal to the error probability and we have

$$E[d(X,\hat{X})] = P(X \ne \hat{X}) = q \le D.$$

By differentiating with respect to q, we can show that $f(q)$ is increasing for $q \le (m-1)/m$. Since $q \le D \le (m-1)/m$, $f(q)$ is maximized when $q = D$. Thus, lower bound of $I(X;\hat{X})$ is

$$I(X;\hat{X}) \ge \log m - [H_B(D) + D \log(m-1)].$$

Achievability: We now show that this lower bound is achievable. To achieve the lower bound, $H(X|\hat{X})$ should be maximized. As derived above, this maximum occurs when $P(X \ne \hat{X}) = D$. Under this condition, the $H(X|\hat{X})$ is maximized when the uncertainty of X given an error is maximized, meaning that, given an error, X is uniformly distributed among the $m-1$ symbols. This leads us to consider the conditional probability,

$$p(x|\hat{x}) = \begin{cases} \frac{D}{m-1} & \text{if } x \ne \hat{x} \\ 1 - D & \text{if } x = \hat{x} \end{cases}$$

For such $p(x|\hat{x})$, mutual information is calculated as:

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= \log m + (m-1) \cdot \frac{D}{m-1} \cdot \log \frac{D}{m-1} + (1-D) \cdot \log(1-D) \\ &= \log m - H_B(D) - D \log(m-1). \end{aligned}$$

This matches the lower bound derived earlier.

Conclusion: Combining the results, the rate distortion function is

$$R(D) = \begin{cases} \log m - H_B(D) - D \log(m-1) & \text{if } 0 \leq D \leq \frac{m-1}{m} \\ 0 & \text{if } D \geq \frac{m-1}{m} \end{cases}$$

10.17

(a) We can choose R such that $R(D) < R < C$.
Since $R < C$, by channel coding theorem, there exists channel code that maps an index $W \in \{1, 2, ..., 2^{nR}\}$ to X^n and Y^n to $\hat{W}$ such that $P(W^n \neq \hat{W}^n) \to 0$ as $n \to \infty$.

$$\begin{aligned} E[d(V^n, \hat{V}^n)] &= P(W^n \neq \hat{W}^n) E[d(V^n, \hat{V}^n) | W^n \neq \hat{W}^n] \\ &\quad + P(W^n = \hat{W}^n) E[d(V^n, \hat{V}^n) | W^n = \hat{W}^n] \\ &= E[d(V^n, \hat{V}^n) | W^n = \hat{W}^n] \end{aligned}$$

Since $R > R(D)$, by rate distortion theorem, there exists a source code that maps V^n to index W and the index W to $\hat{V}^n$ such that $E[d(V^n, \hat{V}^n) | W^n = \hat{W}^n] \to D$ as $n \to \infty$.
Thus, there exists encoder and decoder pair that achieve $E[d(V^n, \hat{V}^n)] \to D$ as $n \to \infty$.

(b) We can expand $I(V^n; \hat{V}^n)$ as follows:

$$\begin{aligned}
I(V^n; \hat{V}^n) &= H(V^n) - H(V^n|\hat{V}^n) \\
&= \sum_{i=1}^{n} H(V_i) - H(V^n|\hat{V}^n) \\
&= \sum_{i=1}^{n} H(V_i) - \sum_{i=1}^{n} H(V_i|V_{i-1}, ..., V_i, \hat{V}^n) \\
&\geq \sum_{i=1}^{n} H(V_i) - \sum_{i=1}^{n} H(V_i|\hat{V}_i) \quad \textbf{(conditioning reduces entropy)} \\
&= \sum_{i=1}^{n} I(V_i, \hat{V}_i) \\
&\geq \sum_{i=1}^{n} R(E[d(V_i, \hat{V}_i)]) \quad \boldsymbol{(R(E[d(V_i, \hat{V}_i)]) = \min I(V_i, \hat{V}_i))} \\
&= n \cdot \frac{1}{n} \sum_{i=1}^{n} R(E[d(V_i, \hat{V}_i)]) \\
&\geq nR(\frac{1}{n} \sum_{i=1}^{n} E[d(V_i, \hat{V}_i)]) \quad \textbf{(by convexity of } R(D)\textbf{)} \\
&= nR(E[d(V^n, \hat{V}^n)]) \\
&= nR(D)
\end{aligned}$$

Thus,

$$nR(D) \leq I(V^n; \hat{V}^n) \overset{(a)}{\leq} I(X^n; Y^n) \overset{(b)}{\leq} nC$$

where (a) follows from data processing inequality and (b) follows from a property of discrete memoryless channel.

$\therefore R(D) \leq C$.

10.15

(a) Decreasing in R.
Consider two rates R_1, R_2 such that $R_1 < R_2$. Let's define a set of feasible $p(\hat{x}|x)$ for each rate as:

$$\mathcal{A}(R) = \{p(\hat{x}|x) : I(X; \hat{X}) \leq R\}.$$

Since $R_1 < R_2$, $\mathcal{A}(R_1) \subseteq \mathcal{A}(R_2)$. Thus,

$$\min_{p \in \mathcal{A}(R_1)} E[d(X, \hat{X})] \geq \min_{p \in \mathcal{A}(R_2)} E[d(X, \hat{X})]$$

Therefore, $D(R)$ is decreasing in R.

(b) Convex in R.
Consider two rates R_1, R_2 and $\lambda \in (0,1)$.
Let $p_1(\hat{x}|x)$ and $p_2(\hat{x}|x)$ be the distributions that achieve the minimum distortion for R_1 and R_2, respectively. That is,

$$I_{p_1}(X;\hat{X}) \le R_1, \quad E_{p_1}[d(X,\hat{X})] = D(R_1)$$
$$I_{p_2}(X;\hat{X}) \le R_2, \quad E_{p_2}[d(X,\hat{X})] = D(R_2)$$

Construct a mixture distribution $p_\lambda(\hat{x}|x) = \lambda p_1(\hat{x}|x) + (1-\lambda)p_2(\hat{x}|x)$. Since mutual information $I(X;\hat{X})$ is convex with respect to $p(\hat{x}|x)$,

$$I_{p_\lambda}(X;\hat{X}) \le \lambda I_{p_1}(X;\hat{X}) + (1-\lambda) I_{p_2}(X;\hat{X}) \le \lambda R_1 + (1-\lambda)R_2.$$

Thus, $p_\lambda(\hat{x}|x)$ satisfies the constraint $I(X;\hat{X}) \le \lambda R_1 + (1-\lambda)R_2$ and hence,

$$\begin{aligned} D(\lambda R_1 + (1-\lambda)R_2) &= \min_{p(\hat{x}|x): I(X;\hat{X}) \le \lambda R_1 + (1-\lambda)R_2} E[d(X,\hat{X})] \\ &\le E_{p_\lambda}[d(X,\hat{X})] \\ &\overset{(a)}{=} \lambda E_{p_1}[d(X,\hat{X})] + (1-\lambda) E_{p_2}[d(X,\hat{X})] \\ &= \lambda D(R_1) + (1-\lambda) D(R_2) \end{aligned}$$

where (a) follows from linearity of expectation.
Therefore, $D(R)$ is convex in R.

(c) (a) follows from the definition of distortion for sequence,
(b) follows from linearity of expectation,
(c) follows from the definition of distortion rate function,
(d) follows from convexity of $D(R)$,
(e) follows from the fact that

$$\begin{aligned} I(X^n;\hat{X}^n) &= H(X^n) - H(X^n|\hat{X}^n) \\ &= \sum_{i=1}^n H(X_i) - H(X^n|\hat{X}^n) \\ &= \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|X_{i-1},...,X_i,\hat{X}^n) \\ &\ge \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\hat{X}_i) \\ &= \sum_{i=1}^n I(X_i;\hat{X}_i) \end{aligned}$$

and the decreasing property of $D(R)$,
(f) follows from $(*)$ data processing inequality and the decreasing property of $D(R)$.

$$I(X^n;\hat{X}^n) \overset{(*)}{\le} I(i(X^n);\hat{X}^n) \le H(i(X^n)) \le \log|i(X^n)| = nR$$