EE 4TM4: Digital Communications II

Channel Capacity

I. Channel Coding Theorem

Definition 1: A rate R is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)}(\mathcal{C}) = 0$. The capacity C of a DMC is the supremum over all achievable rates.

Theorem 1: The capacity of the discrete memoryless channel $p(y|x)$ is given by the information capacity formula

$$C = \max_{p(x)} I(X;Y).$$

Lemma 1: Let $(X,Y) \sim p_{X,Y}(x,y)$ and $p(x^n,y^n) = \prod_{i=1}^n p_X(x_i)p_Y(y_i)$. Then

$$P\{(X^n,Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} \le 2^{-n(1-\epsilon)I(X;Y)};$$

moreover,

$$P\{(X^n,Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} \ge (1-\epsilon)2^{-n(1+\epsilon)I(X;Y)}$$

for n sufficiently large.

Proof: Note that

$$P\{(X^n,Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} = \sum_{(x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)} p(x^n)p(y^n). \tag{1}$$

For $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$ and $y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$,

$$2^{-(1+\epsilon)H(X)} \le p(x^n) \le 2^{-n(1-\epsilon)H(X)}, \tag{2}$$

$$2^{-(1+\epsilon)H(Y)} \le p(y^n) \le 2^{-n(1-\epsilon)H(Y)}. \tag{3}$$

Recall that

$$|\mathcal{T}_\epsilon^{(n)}(X,Y)| \le 2^{n(1+\epsilon)H(X,Y)}; \tag{4}$$

moreover, we have

$$|\mathcal{T}_\epsilon^{(n)}(X,Y)| \ge (1-\epsilon)2^{n(1-\epsilon)H(X,Y)} \tag{5}$$

for n sufficiently large. Substituting (2)–(5) into (1) completes the proof. ■

Achievability. For every rate $R < C = \max_{p(x)} I(X;Y)$, there exists a sequence of $(2^{nR}, n)$ codes with average probability of error $P_e^{(n)}$ that tends to zero as $n \to \infty$. The proof of achievability uses random coding and joint typicality decoding.

Random codebook generation. We use random coding. Fix the pmf $p(x)$ that attains the information capacity C. Randomly and independently generate 2^{nR} sequences $x^n(m)$, $m \in [1 : 2^{nR}]$, each according to $p(x^n) = \prod_1^n p_X(x_i)$. The generated sequences constitute the codebook $\mathcal{C}$.

Encoding. To send a message $m \in [1 : 2^{nR}]$, transmit $x^n(m)$.

Decoding. We use joint typicality decoding. Let y^n be the received sequence. The receiver declares that $\hat{m} \in [1 : 2^{nR}]$ is sent if it is the unique message such that $(x^n(\hat{m}), y^n) \in \mathcal{T}_\epsilon^{(n)}$; otherwise-if there is none or more than one such message-it declares an error e.

Analysis of the probability of error. Assuming that message m is sent, the decoder makes an error if $(x^n(m), y^n) \notin \mathcal{T}_\epsilon^{(n)}$ or if there is another message $m' \neq m$ such that $(x^n(m'), y^n) \in \mathcal{T}_\epsilon^{(n)}$.

Consider the probability of error averaged over M and over all codebooks

$$\begin{aligned} P(\mathcal{E}) &= E_{\mathcal{C}}(P_e^{(n)}) \\ &= E_{\mathcal{C}}(2^{-nR} \sum_{m=1}^{2^{nR}} \lambda_m(\mathcal{C})) \\ &= 2^{-nR} \sum_{m=1}^{2^{nR}} E_{\mathcal{C}}(\lambda_m(\mathcal{C})) \\ &= E_{\mathcal{C}}(\lambda_1(\mathcal{C})) = P(\mathcal{E}|M = 1). \end{aligned}$$

Thus we assume without loss of generality that $M = 1$ is sent. For brevity, we do not explicitly condition on the event $\{M = 1\}$ in probability expressions whenever it is clear from the context.

The decoder makes an error iff one or both of the following events occur:

$$\begin{aligned} \mathcal{E}_1 &= \{(X^n(1), Y^n) \in \mathcal{T}_\epsilon^{(n)}\}, \\ \mathcal{E}_2 &= \{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)} \text{ for some } m \neq 1\}. \end{aligned}$$

Thus, by the union of events bound,

$$P(\mathcal{E}) = P(\mathcal{E}_1 \cup \mathcal{E}_2) \leq P(\mathcal{E}_1) + P(\mathcal{E}_2).$$

We now bound each term. By the weak law of large numbers, the first term $P(\mathcal{E}_1)$ tends to zero as $n \to \infty$. For the second term, since for $m \neq 1$,

$$(X^n(m), X^n(1), Y^n) \sim \prod_{i=1}^n p_X(x_i(m)) p_{X,Y}(x_i(1), y_i),$$

we have $(X^n(m), Y^n) \sim \prod_{i=1}^n p_X(x_i(m)) p_Y(y_i)$. Thus, by Lemma 1,

$$P\{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)}\} \leq 2^{-n(I(X;Y)-\delta(\epsilon))} = 2^{-n(C-\delta(\epsilon))}.$$

Again by the union of events bound,

$$P(\mathcal{E}_2) \leq \sum_{m=2}^{2^{nR}} P\{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)}\} \leq \sum_{m=2}^{2^{nR}} 2^{-n(C-\delta(\epsilon))} \leq 2^{-n(C-R-\delta(\epsilon))},$$

which tends to zero as $n \to \infty$ if $R < C - \delta(\epsilon)$.

To complete the proof, note that since the probability of error averaged over the codebooks, $P(\mathcal{E})$, tends to zero as $n \to \infty$, there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_\epsilon^{(n)} = 0$. This proves that $R < I(X;Y) = C$ is achievable.

Remark: By the Markov inequality, the probability of error for a random codebook, that is, a codebook consisting of random sequences $X^n(m)$, $m \in [1 : 2^{nR}]$, tends to zero as $n \to \infty$ in probability. Hence, most codebooks are good in terms of the error probability.

Remark: The capacity with the maximal probability of error $\lambda^* = \max_m \lambda_m$ is equal to that with the average probability of error $P_e^{(n)}$. This can be shown by throwing away the worst half of the codewords (in terms of error probability) from each of the sequence of $(2^{nR}, n)$ codes that achieve R. The maximal probability of error for the codes with the remaining codewords is at most $2P_e^{(n)}$, which again tends to zero as $n \to \infty$. As we shall see, the capacity with maximal probability is not always the same as that with the average probability of error for multiuser channels.

Remark: The capacity for constant composition codes is the same due to the fact that there are only polynomial number of types. First make sure that the maximal error is small, then reduce to a constant composition code (otherwise the resulting constant composition code might have poor error probability).

Converse. For every sequence of $(2^{nR}, n)$ codes with $P_e^{(n)}$ that tends to zero as $n \to \infty$, the rate $R \le C = \max_{p(x)} I(X;Y)$. The proof of the converse uses Fano's inequality and basic properties of entropy and mutual information.

Every $(2^{nR}, n)$ code induces a joint pmf on (M, X^n, Y^n) of the form

$$p(m, x^n, y^n) = 2^{-nR} p(x^n|m) \prod_{i=1}^{n} p_{Y|X}(y_i|x_i).$$

By Fano's inequality,

$$H(M|Y^n) \le 1 + P_e^{(n)} nR = n\epsilon_n,$$

where ϵ_n tends to zero as $n \to \infty$ by the assumption that $\lim_{n\to\infty} P_e^{(n)} = 0$. Now consider

$$\begin{aligned}
nR &= H(M) \\
&= I(M;Y^n) + H(M|Y^n) \\
&\le I(M;Y^n) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(M;Y_i|Y^{i-1}) + n\epsilon_n \\
&\le \sum_{i=1}^{n} I(M, Y^{i-1}; Y_i) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(X_i, M, Y^{i-1}; Y_i) + n\epsilon_n
\end{aligned}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n$$
$$\leq nC + n\epsilon.$$

Remark: Discuss the condition under which the inequalities in the converse become equalities, which imposes constraints on the optimal codebook.

Channel coding with input cost

Consider a DMC $p(y|x)$. Suppose that there is a nonnegative cost $b(x)$ associated with each input symbol $x \in \mathcal{X}$. Without loss of generality, we assume that there exists a zero cost symbol $x_0 \in \mathcal{X}$, i.e., $b(x_0) = 0$. We further assume that an average cost constraint is imposed on each codeword, i.e.,

$$\sum_{i=1}^{n} b(x_i(m)) \leq nB, \quad m \in [1 : 2^{nR}].$$

Now, defining the channel capacity of the DMC with cost constraint B, or the capacity-cost function, $C(B)$ in a similar manner to capacity without cost constraint, we can establish the following extension of the channel coding theorem.

Theorem 2: The capacity of the DMC $p(y|x)$ with cost constraint B is

$$C(B) = \max_{p(x): E[b(X)] \leq B} I(X; Y).$$

Note that $C(B)$ is nondecreasing, concave, and continuous in B.

Proof of achievability. Fix the pmf $p(x)$ that attains $C(B/(1+\epsilon))$. Randomly and independently generate 2^{nR} sequences $x^n(m)$, $m \in [1 : 2^{nR}]$, each according to $\prod_{i=1}^{n} p_X(x_i)$. To send message m, the encoder transmits $x^n(m)$ if $x^n(m) \in \mathcal{T}_\epsilon^{(n)}$, and consequently, by the typicality average lemma, the sequence satisfies the cost constraint $\sum_{i=1}^{n} b(x_i(m)) \leq nB$. Otherwise, it transmits $(x_0, \cdots, x_0)$. The analysis of the average probability of error for joint typicality decoding follows similar lines to the case without cost constraint. Assume $M = 1$. For the first probability of error event,

$$\begin{aligned}
P(\mathcal{E}_1) &= P\{(X^n(1), Y^n) \in \mathcal{T}_\epsilon^{(n)}\} \\
&= P\{X^n(1) \in \mathcal{T}_\epsilon^{(n)}, (X^n(1), Y^n) \notin \mathcal{T}_\epsilon^{(n)}\} + P\{X^n(1) \notin \mathcal{T}_\epsilon^{(n)}, (X^n(1), Y^n) \notin \mathcal{T}_\epsilon^{(n)}\} \\
&\leq \sum_{x^n \in \mathcal{T}_\epsilon^{(n)}} \prod_{i=1}^{n} p_X(x_i) \sum_{y^n \notin \mathcal{T}_\epsilon^{(n)}(Y|x^n)} \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) + P\{X^n(1) \notin \mathcal{T}_\epsilon^{(n)}\} \\
&\leq \sum_{(x^n, y^n) \notin \mathcal{T}_\epsilon^{(n)}} \prod_{i=1}^{n} p_X(x_i) p_{Y|X}(y_i|x_i) + P\{X^n(1) \notin \mathcal{T}_\epsilon^{(n)}\}.
\end{aligned}$$

By the LLN for each term, this tends to zero as $n \to \infty$. The probability of error for the second event, $P(\mathcal{E}_2)$, is upper bounded in exactly the same manner as when there is no cost constraint. Hence, every rate $R < I(X; Y) = C(B/(1+\epsilon))$ is achievable. Finally, by the continuity of $C(B)$ in B, $C(B/(1+\epsilon))$ converges to $C(B)$ as $\epsilon \to 0$, which implies the achievability of every rate $R < C(B)$.

Proof of the converse. Consider a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$ such that for every n, the cost constraint $\sum_{i=1} b(x_i(m)) \le nB$ is satisfied for every $m \in [1 : 2^{nR}]$ and thus $\sum_{i=1}^n E_M[b(X_i(M))] \le nB$. As before, by Fano's inequality and the data processing inequality,

$$\begin{aligned} nR &\le \sum_{i=1}^n I(X_i; Y_i) + n\epsilon_n \\ &\le \sum_{i=1}^n C(E[b(X_i)]) + n\epsilon_n \\ &\le nC\left(\frac{1}{n}\sum_{i=1}^n E[b(X_i)]\right) + n\epsilon_n \\ &\le nC(B) + n\epsilon. \end{aligned}$$

Remark: The converse proof shows that the capacity is the same if we impose a weaker average cost constraint $2^{-nR}\sum_m \sum_{i=1}^n b(x_i(m)) \le nB$.

II. Channel with State

Channel model: $p(Y(t)|X^t, S^t) = p(Y(t)|X(t), S(t))$. The state process $\{S(t)\}_{t=-\infty}^{\infty}$ is stationary and memoryless[1] with marginal distribution $\pi(s)$ over state space $\mathcal{S}$, and is unaffected by the input and output processes.

1) No Knowledge of Channel State Information:

   The channel capacity is given by

   $$C = \max_{p(x)} I(X; Y).$$

   Here the channel transition probability is given by

   $$p(y|x) = \sum_{s\in\mathcal{S}} \pi(s) p(y|x, s).$$

2) Perfect Channel State Information at Receiver:

   In this case, we can view $\{Y(t), S(t)\}$ as the channel output, and thus the capacity formula is given by

   $$C = \max_{p(x)} I(X; Y, S).$$

   Since the transmitter does not have the channel state information, it follows that $I(X; S) = 0$, which yields

   $$I(X; Y, S) = I(X; Y|S) + I(X; S) = I(X; Y|S).$$

   Thus the capacity formula can be simplified to

   $$C = \max_{p(x)} I(X; Y|S).$$

3) Perfect Channel State Information at both Transmitter and Receiver:

[1]This condition can be relaxed.

The channel capacity is given by

$$C = \sum_{s \in \mathcal{S}} \pi(s) \max_{p(x|s)} I(X;Y|S=s).$$

This capacity has an intuitive interpretation. Since both the transmitter and the receiver know the state realization, they can use the multiplexing technique to decompose the channel to $|\mathcal{S}|$ memoryless channels, each corresponding to a state $s \in \mathcal{S}$. For each state s, the optimal coding scheme is used to achieve the memoryless channel capacity with respect to state s, i.e.,

$$C_s = \max_{p(x|s)} I(X;Y|S=s).$$

Since the state process is stationary and ergodic, given n time slots, the number of times that the state has realization s is approximately $n\pi(s)$ as n is large enough. Therefore, the average rate is

$$\sum_{s \in \mathcal{S}} \pi(s) C_s = \sum_{s \in \mathcal{S}} \pi(s) \max_{p(x|s)} I(X;Y|S=s).$$

Note: Although the multiplexing method gives us a simple way to achieve that channel capacity, it is not optimal in the sense of error exponent since the dependency in the state process is not fully exploited. Furthermore, this multiplexing method requires extremely long block coding when the state space is large, so it may cause long delay in decoding.

REFERENCES

[1] T. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition. Hoboken, NJ: Wiley, 2006.

[2] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2011.