ECE253/CSE208 Introduction to Information Theory

Lecture 11: Channel Coding Theorem

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 7 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*

Communication Diagram

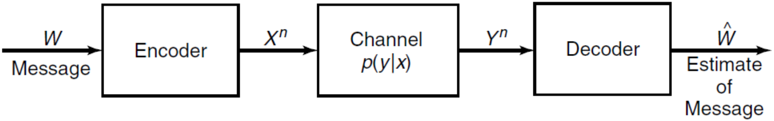

Figure: A diagram showing how a message is communicated through a noisy channel.

- Essentially, the communication system represents a Markov chain:

$$W \to X^n \to Y^n \to \hat{W}.$$

- Here, the encoder/decoder block represents a joint source-channel encoder/decoder.

Shannon's Second Theorem

- Reliable (virtually error-free) communication is possible at rates up to the capacity.
- Channel capacity is the sharp threshold between reliable and unreliable communication.

Theorem (Channel Coding Theorem)

For a DMC, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda_{\max}^{(n)} \to 0$ as $n \to \infty$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda_{\max}^{(n)} \to 0$ must have $R \leq C$.

Let $P_e^{(n)}$ denote the average probability of error, and A be a given finite non-negative constant. The weak and strong versions of the converse statement are given as follows.

- *Weak converse:* $P_e^{(n)} \geq 1 - \frac{1}{nR} - \frac{C}{R} \implies$ If $R > C$, $P_e^{(n)}$ is bounded away from zero as $n \to \infty$.
- *Strong converse:* $P_e^{(n)} \geq 1 - \frac{4A}{n(R-C)^2} - e^{\frac{-n(R-C)}{2}} \implies$ If $R > C$, $P_e^{(n)} \xrightarrow{n \to \infty} 1$.

Discrete Channel and Its Extension

A few definitions are needed for the proof of the channel coding theorem.

Definition

A discrete channel, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of two finite sets $\mathcal{X}, \mathcal{Y}$ and a collection of probability mass functions $p(y|x)$. Assume $p(y|x) \geq 0$ for all (x, y). For all x, $\sum_y p(y|x) = 1$. Note that (x, y) is the input-output pair of the channel.

Definition

The n-th extension of the DMC is $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$ for $k = 1, 2, \ldots, n$.

For a channel without feedback, we have

$$p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x_{k-1}) \Rightarrow p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i).$$

(M, n) Code and Code Rate

Definition ((M, n) code)

An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. Message $W \in \{1, 2, \ldots, M\} \triangleq \mathcal{M}$, where M is the size of the message set.
2. An encoding function: $X^n : \mathcal{M} \to \mathcal{X}^n$ yields the codebook $\mathcal{C} = [x^n(1), \ldots, x^n(M)]$.
3. A deterministic decoding function: $g : \mathcal{Y}^n \to \mathcal{M}$ yields an estimate $\hat{W}$.

Definition

The rate R of an (M, n) code is $R = \frac{\log M}{n}$ bits per transmission.

- **Code rate R is the number of info bits conveyed per channel use.** If we only consider channel coding, then for every $k \triangleq \log M$ bits of useful information, the coder generates a total of n bits of data, of which $n - k$ are redundant for error detection/correction. Hence, the rate R quantifies the coder's efficiency.
- For notational simplicity, we write $(2^{nR}, n)$ codes to mean $(\lceil 2^{nR} \rceil, n)$ codes.

Probability of Error

Definition

- The **conditional** probability of error is

$$\boxed{\lambda_i := \Pr\left(g(Y^n) \neq i \mid X^n = x^n(i)\right) = \sum_{y^n} p(y^n|x^n(i)) \times \mathbb{1}(g(y^n) \neq i)}$$

- The **maximum** probability of error $\lambda_{\max}^{(n)}$ for an (M, n) code is

$$\boxed{\lambda_{\max}^{(n)} = \max_{i \in \{1,\dots,M\}} \lambda_i.}$$

- The **average** probability of error is

$$\boxed{P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.}$$

Clearly, we have $P_e^{(n)} \leq \lambda_{\max}^{(n)}$. If the message W is chosen uniformly over $\mathcal{M}$ and $X^n = x^n(w)$, then $P_e^{(n)} = \Pr(W \neq g(Y^n))$.

Achievable Rate

Definition

A rate R is said to be achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that $\lambda_{\max}^{(n)} \xrightarrow{n \to \infty} 0$.

Definition

The capacity of a channel is the supremum[a] of all achievable rates.

[a] In terms of sets, the *maximum* is the largest member of the set while the *supremum* is the smallest upper bound of the set. supremum = maximum for compact sets.

Joint Typicality

Roughly speaking, we decode as $g(Y^n) \mapsto w$, if $X^n(w)$ is *jointly typical* with Y^n.
Recall that a typical sequence has its empirical entropy ϵ-close to the true entropy $H(X)$.

$$A_\epsilon^{(n)} = \Bigg\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|-\frac{1}{n}\log p(X^n) - H(X)\right| < \epsilon, \quad \left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| < \epsilon,$$
$$\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| < \epsilon\Bigg\}.$$

Theorem (Joint AEP)

Let (X^n, Y^n) *be i.i.d.* $\sim p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$*. Then,*

1. $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \xrightarrow{n\to\infty} 1.$
2. $|A_\epsilon^{(n)}| \le 2^{n(H(X,Y)+\epsilon)}.$
3. *If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$*, then*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \le \Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \le 2^{-n(I(X;Y)-3\epsilon)},$$

where the upper bound holds for n sufficiently large.

Intuitive Proof of the Theorem

- Typical sets $|X^n| \approx 2^{nH(X)}$ and $|Y^n| \approx 2^{nH(Y)}$.
- Not all pair of typical X^n and typical Y^n are jointly typical: only about $2^{nH(X,Y)}$ paris are joint typical.
- The probability of any randomly chosen pair is jointly typical is about

$$\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X;Y)}.$$

- This implies that there are about $2^{nI(X;Y)}$ distinguishable input signals X^n.
- So if the number of possible input codewords is 2^{nR} with $R = I(X;Y) - \epsilon$, then $P_e^{(n)} = 2^{nR} \times 2^{-nI(X;Y)} \leq 2^{-n\epsilon} \to 0$ as $n \to \infty$.

Intuitive Proof: Sphere Packing

- For large block lengths, every channel looks like the noisy typewriter channel.
- For each (typical) input X sequence, there are about $2^{nH(Y|X)}$ possible Y sequences (all of them equally likely). we have about $2^{nH(Y)}$ typical Y sequences.
- Hence, the total number of disjoint sets we can afford is $\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)}$ and C is no greater than $I(X;Y)$ (maximized over $p(x)$).

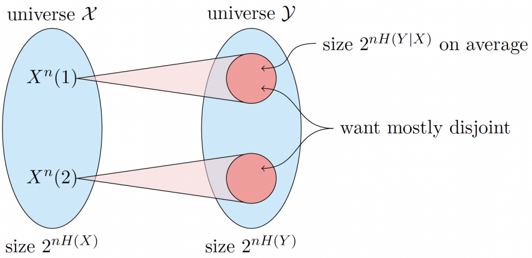

Figure: How many distinguishable input sequences X^n can produce disjoint sequences at the output? Figure credit to V. Guruswami's lecture note.

Proof Outline

1. At the transmitter, use **random coding**.

2. $W \in \{1, 2, \ldots, 2^{nR}\}$ has a uniform distribution.

3. At the receiver, use **jointly typical decoding** for Y^n to find $X^n(w)$.
   We will bound two types of error:
   - Type-1 error: $X^n(w)$ is not jointly typical with Y^n; and
   - Type-2 error: find a sequence $\tilde{X}^n(\hat{w})$ is jointly typical with Y^n, but $\hat{w} \neq w$.

4. Use joint AEP to prove achievability (direct part) and Fano's inequality for the converse statement.

Proof of the Channel Coding Theorem

On the sender side, do the following:

1. Randomly generate a $\left(2^{nR}, n\right)$ code according to a fixed $p(x)$. Specifically, we generate 2^{nR} codewords independently according to the distribution $p\left(x^n\right) = \prod_{i=1}^{n} p\left(x_i\right)$. Collect codewords as the rows of the codebook:

$$\mathcal{C} = \left[\begin{array}{cccc} x_1(1) & x_2(1) & \ldots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1\left(2^{nR}\right) & x_2\left(2^{nR}\right) & \cdots & x_n\left(2^{nR}\right) \end{array} \right].$$

   The codebook $\mathcal{C}$ is known to both the encoder and the decoder.

   Each entry is i.i.d. $\sim p(x)$. Thus, the probability of a particular code $\mathcal{C}$ is given by

$$\Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p\left(x_i(w)\right).$$

2. Uniformly choose a message W: $\Pr(W = w) = 2^{-nR},\ w = 1, 2, \ldots, 2^{nR}$.

Proof of the Channel Coding Theorem (cont'd)

On the receiver side, do the following:

1. Obtain a sequence Y^n according to $\Pr(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))$.
2. Guess which message was sent. For the *jointly typical decoding*, the receiver declares:
   - **index $\hat{W}$ was sent** if $(X^n(\hat{W}), Y^n)$ is jointly typical, and there is no other message W' such that $(X^n(W'), Y^n)$ is jointly typical.
   - **an error** if no such $\hat{W}$ or more than one such.
3. Calculate the probability of errors. Let $\mathcal{E} = \left\{\hat{W}(Y^n) \neq W\right\}$ denote the error event.

$$P_e^{(n)} = \Pr(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) = \Pr(\mathcal{E} \mid W = 1).$$

Note that this is the probability of error averaged over all codebooks and codewords.

Proof of the Channel Coding Theorem (cont'd)

4. For $i \in \{1, 2 \ldots, 2^{nR}\}$, let $E_i \triangleq \left\{(X^n(i), Y^n) \in A_\epsilon^{(n)}\right\}$ denote the event that the i-th codeword and Y^n are jointly typical. WLOG, assume that Y^n is the received sequence by sending $X^n(1)$ over the channel.

$$\Pr(\mathcal{E}|W=1) = \Pr\left(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W=1\right) \tag{1}$$

$$\leq \underbrace{\Pr\left(E_1^c|W=1\right)}_{\text{type-I error}} + \sum_{i=2}^{2^{nR}} \underbrace{\Pr\left(E_i|W=1\right)}_{\text{type-II error}} \tag{2}$$

By the joint AEP, we have

- For the type-I error, $\Pr\left(E_1^c|W=1\right) \leq \epsilon$ for n sufficiently large.
- For the type-II error, $\Pr\left(E_i|W=1\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$, $\forall i \neq 1$. Note that for any $i \neq 1$, Y^n and $X^n(i)$ are independent.

Proof of the Channel Coding Theorem (cont'd)

Therefore, we have

$$
\begin{aligned}
\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W=1) &\leq \Pr\left(E_1^c|W=1\right) + \sum_{i=2}^{2^{nR}} \Pr\left(E_i|W=1\right) && (3)\\
&\leq \epsilon + (2^{nR}-1) \times 2^{-n(I(X;Y)-3\epsilon)} && (4)\\
&\leq \epsilon + 2^{-n(I(X;Y)-3\epsilon-R)} && (5)\\
&\leq 2\epsilon, && (6)
\end{aligned}
$$

if n is sufficiently large and $R < I(X;Y) - 3\epsilon$.

Hence, if $R < I(X;Y)$, we can choose ϵ and n so that the average probability of error is less than 2ϵ.

Proof of the Channel Coding Theorem (cont'd)

We can strengthen the conclusion by a series of code selections.

1. In the proof, set $p(x) = p^*(x)$, which is the optimal input distribution achieving the capacity. Then the condition $R < I(X;Y)$ becomes $R < C$.
2. Get rid of the averaging over codebooks. Since the average probability of error over codebooks is less than 2ϵ, there exists at least one codebook $\mathcal{C}^*$ such that

$$\Pr(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*) \leq 2\epsilon. \tag{7}$$

   We can find $\mathcal{C}^*$ by exhaustive search over a total of $|\mathcal{X}|^{Mn}$ possible codebooks.
3. Relate $P_e^{(n)}$ to $\lambda_{\max}^{(n)}$: Discard the worst half of the codewords in $\mathcal{C}^*$.
   Due to (7), we know that at least half the $\lambda_i(\mathcal{C}^*)$ are less than 4ϵ. If we keep this half the codewords and discard the remaining half, we get $\lambda_{\max}^{(n)} \leq 4\epsilon$ while the rate changes from R to $R - \frac{1}{n}$ (negligible for large n).

Zero-error Codes

The outline of the proof of the converse is most clearly motivated by going through the argument when absolutely no errors are allowed. We now prove that $P_e^{(n)} = 0$ implies that $R \leq C$.

Proof:

$$nR = H(W) = \underbrace{H(W|Y^n)}_{=0} + I(W;Y^n)$$

$$= I(W;Y^n) \leq I(X^n;Y^n) \leq \sum_{i=1}^{n} I(X_i;Y_i) \leq nC$$

Lemma (Fano's inequality)

For a DMC with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have $H(W|\hat{W}) \leq 1 + P_e^{(n)} nR$.

Proof of the Converse Statement

Lemma

For a DMC of capacity C, we have $I(X^n; Y^n) \leq nC$ for all $p(x^n)$.

Proof:

$$\begin{aligned} I(X^n; Y^n) = H(Y^n) - H(Y^n | X^n) &= H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y_1, \ldots, Y_{i-1}, X^n) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \leq nC \end{aligned}$$

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq H(W|\hat{W}) + I(X^n;Y^n) \leq 1 + P_e^{(n)} nR + nC \implies \end{aligned}$$

$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR} \implies$ If $R > C$, then $P_e^{(n)}$ is bounded away from 0 as $n \to \infty$.

Feedback Capacity

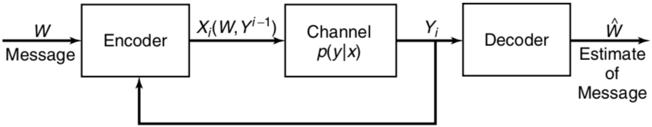

Figure: Discrete memoryless channel (DMC) with feedback.

For a DMC, feedback may help simplify encoding and decoding (e.g., for BEC), but will not increase the channel capacity.

Theorem (Feedback does not increase the channel capacity for a DMC)

For a DMC, $C_{FB} = C = \max_{p(x)} I(X;Y)$.

Proof: Clearly, $C_{FB} \geq C$. We need to show $C_{FB} \leq C$.

$$
\begin{aligned}
nR = H(W|\hat{W}) + I(W;\hat{W}) &\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \\
&\leq 1 + P_e^{(n)} nR + \boxed{I(W;Y^n)} \quad \text{[by DPI]}
\end{aligned}
$$

Proof (cont'd)

$$\boxed{I(W;Y^n)} = H(Y^n) - H(Y^n|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},W)$$

$$= H(Y^n) - \sum_{i=1}^{n} H\left(Y_i|Y_1,\ldots,Y_{i-1},W,X_i(W,Y^{i-1})\right) \quad \text{[due to feedback]}$$

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i;Y_i) \leq nC \implies$$

$$nR \leq 1 + P_e^{(n)} nR + nC \implies \boxed{R \leq \frac{1}{n} + P_e^{(n)} R + C.}$$

Finally, taking $n \to \infty$ and $P_e^{(n)} \to 0$, we get $R \leq C$.

Remark

- $C_{\text{FB}} = C$: **A higher rate with feedback cannot be achieved for a DMC**.
- The availability of feedback often makes coding simpler.
- In general, if the channel has memory, feedback can increase the capacity.

Communication above Capacity[1]

Theorem (Communication with error)

If a probability of bit error P_b is acceptable, rates up to $R(P_b)$ are achievable, where

$$R(P_b) = \frac{C}{1 - H(P_b)}.$$

For any P_b, rates greater than $R(P_b)$ are not achievable.

[1] Page 162 of the book "Information Theory, Inference and Learning Algorithms" by David J. MacKay

Thank You!

Email: `<zhangy@ucsc.edu>`

Homepage: `https://people.ucsc.edu/~yzhan419/`