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*

Shannon's Second Theorem

Channel capacity is the sharp threshold between reliable and unreliable communication. Informal statement: For a DMC,

1. All rates below capacity $R < C$ are achievable.
2. Conversely, $(2^{nR}, n)$ code with probability of error $\lambda^{(n)} \xrightarrow{n\to\infty} 0$ must have $R \leq C$.

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 $\lambda^{(n)} \to 0$ as $n \to \infty$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ 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 Channels

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

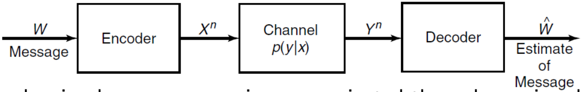

Figure: A diagram showing how a message is communicated through a noisy channel, which essentially represents a Markov chain: $W \to X^n \to Y^n \to \hat{W}$.

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 memoryless 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.

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

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

  where $\mathbb{1}(\cdot)$ is the indicator function.

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

$$\boxed{\lambda^{(n)} = \max_{i \in \{1,\ldots,M\}} \lambda_i.}$$

- The **average** probability of error is

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

Clearly, we have $P_e^{(n)} \leq \lambda^{(n)}$. 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 $\lambda^{(n)} \xrightarrow{n \to \infty} 0$.

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.

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 $H(X)$.

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

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. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, *then*

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

where the upper bound holds for n sufficiently large.

Preview 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 $2^{nH(X,Y)}$.
- 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 signals X^n.

Proof Outline

1. Use random coding scheme for encoding.

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

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

4. Properties of the joint AEP is used to prove achievability. To prove the converse statement, we use Fano's inequality that relates $P_e^{(n)}$ with $H(W|\hat{W})$.

Proof of the Channel Coding Theorem

On the sender side, do the following:

1. Randomly generate a $\left(2^{nR}, n\right)$ code $\sim$ 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:

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

   Each entry is i.i.d. $\sim p(x)$. Thus, the probability of a particular code $\mathcal{C}$ is $\Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p\left(x_i(w)\right)$.

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 $\Pr\left(y^n|x^n(w)\right) = \prod_{i=1}^n p\left(y_i|x_i(w)\right)$.
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 $\mathcal{E} = \left\{\hat{W}\left(Y^n\right) \neq W\right\}$ denote the error event.

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

Proof of the Channel Coding Theorem (Cont'd)

4. Let $E_i = \left\{(X^n(i), Y^n) \in A_\epsilon^{(n)}\right\}$ for $i \in \{1, 2 \ldots, 2^{nR}\}$, denote the event that the i-th codeword and Y^n are jointly typical. WLOG, Y^n is the result of sending the first codeword $X^n(1)$ over the channel.

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &= \Pr\left(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W=1\right) \\ &\le \Pr\left(E_1^c|W=1\right) + \sum_{i=2}^{2^{nR}} \Pr\left(E_i|W=1\right)\end{aligned}$$

By the joint AEP, we have
$\Pr\left(E_1^c|W=1\right) \to 0 \implies \Pr\left(E_1^c|W=1\right) \le \epsilon$ for n sufficiently large. Since $X^n(1)$ and $X^n(i)$ are independent for any $i \neq 1$, Y^n and $X^n(i)$ are independent too.

Proof of the Channel Coding Theorem (Cont'd)

Therefore, we have

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

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 (averaged over codebooks and codewords) is less than 2ϵ.

Proof of the Channel Coding Theorem (Cont'd)

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

1. Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieves capacity. Then the condition $R < I(X;Y)$ can be replaced by the achievability condition $R < C$.
2. Get rid of the average over codebooks. Since the average probability of error over codebooks is small $(\leq 2\epsilon)$, there exists at least one codebook $\mathcal{C}^*$ such that $\Pr\left(\mathcal{E}|\mathcal{C}^*\right) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i\left(\mathcal{C}^*\right) \leq 2\epsilon$. Determination of $\mathcal{C}^*$ can be achieved by an exhaustive search.
3. Throw away the worst half of the codewords in the best codebook $\mathcal{C}^*$. Since the arithmetic average probability of error $P_e^{(n)}\left(\mathcal{C}^*\right)$ for this code is less then 2ϵ, we have $\Pr\left(\mathcal{E}|\mathcal{C}^*\right) \leq \frac{1}{2^{nR}} \sum \lambda_i\left(\mathcal{C}^*\right) \leq 2\epsilon$ which implies that at least half the indices i and their associated codewords $X^n(i)$ must have conditional probability of error λ_i less than 4ϵ, that is, $\lambda^{(n)} \leq 4\epsilon$. Throwing out half the codewords has changed the rate from R to $R - \dfrac{1}{n}$, which is 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:

$$nR = H(W) = \underbrace{H\left(W|Y^n\right)}_{=0} + I\left(W;Y^n\right)$$
$$= I\left(W;Y^n\right) \leq I\left(X^n;Y^n\right) \leq \sum_{i=1}^{n} I\left(X_i;Y_i\right) \leq nC$$

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 $I\left(X^n; Y^n\right) \leq nC$ for all $p(x^n)$.

Proof:

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

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

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

Thank You!

Email: <zhangy@ucsc.edu>

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