Lecture #10B: Noisy channel coding theorem
Outline of the lecture

- Noisy channel coding theorem
- Converse to noisy channel coding theorem
An \((M, n)\) code for the channel \((X, p(y|x), Y)\) consists of the following:
Definition

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

1) An index set \(\{1, 2, \cdots, M\}\).
2) An encoding function \(X^n : \{1, 2, \cdots, M\} \to X^n\) yielding codewords \(X^n(1), X^n(2), \cdots, X^n(M)\). The set of codewords is referred to as codebook.
Definition

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

1) An index set \(\{1, 2, \cdots, M\}\).
2) An encoding function \(X^n : \{1, 2, \cdots, M\} \to X^n\) yielding codewords \(X^n(1), X^n(2), \cdots, X^n(M)\). The set of codewords is referred to as codebook.
3) A decoding function

\[ g : Y^n \to \{1, 2, \cdots, M\} \]

is a deterministic rule that assigns a guess to each possible received vector.

\[ \lambda_i = Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n} p(y^n | x^n(i)) I(g(y^n) \neq i) \]

be the conditional probability of error given that index \(i\) was sent, where \(I(\cdot)\) is the indicator function.
Definition

Let

\[ \lambda_i = Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n} p(y^n | x^n(i)) I(g(y^n) \neq i) \]

be the conditional probability of error given that index \( i \) was sent, where \( I(\cdot) \) is the indicator function.

The maximal probability of error \( \lambda^n \) for an \((M, n)\) code is defined as

\[ \lambda^n = \max_{i \in 1, 2, \cdots, M} \lambda_i \]

An introduction to Information Theory

Adrish Banerjee
Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh India
Definition

The rate $R$ of an $(M,n)$ code is given by

$$R = \frac{\log M}{n} \text{ bits per transmission}$$

A rate $R$ is said to be achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda(n)$ tends to 0 as $n \to \infty$. 
Definition

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

- A rate $R$ is said to be achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda(n)$ tends to 0 as $n \to \infty$.

- The capacity of a discrete memoryless channel is the supremum of all achievable rates.

Noisy channel coding theorem

- 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 $\lambda(n) \to 0$. 

Adrish Banerjee
Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh India

An introduction to Information Theory
Noisy channel coding theorem

- 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 $\lambda(n) \to 0$

Proof:
- We fix $p(x)$ and generate independently $2^{nR}$ codewords according to the distribution $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- We exhibit the $2^{nR}$ codewords as rows of the matrix

$$C = \begin{bmatrix}
  x_1(1) & x_2(1) & \cdots & x_n(1) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR})
\end{bmatrix}$$
Noisy channel coding theorem

- 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 $\lambda(n) \to 0$

Proof:
- We fix $p(x)$ and generate independently $2^{nR}$ codewords according to the distribution $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- We exhibit the $2^{nR}$ codewords as rows of the matrix

$$C = \begin{bmatrix}
  x_1(1) & x_2(1) & \cdots & x_n(1) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR})
\end{bmatrix}$$

- Probability that we generate a particular code $C$ is

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

Randomly generated code is then revealed to both sender and receiver. Channel transition matrix $p(y/x)$ is known to both sender and receiver.
Noisy channel coding theorem

- Randomly generated code is then revealed to both sender and receiver. Channel transition matrix $p(y|x)$ is known to both sender and receiver.
- A message $W$ is chosen according to a uniform distribution

$$Pr(W = w) = 2^{-nR}, \quad w = 1, 2, \ldots, 2^{nR}$$

- The $w$th codeword $X^n(w)$ corresponding to the $w$th row of $C$ is sent over the channel.
Noisy channel coding theorem

- Randomly generated code is then revealed to both sender and receiver. Channel transition matrix $p(y|x)$ is known to both sender and receiver.
- A message $W$ is chosen according to a uniform distribution
  \[
  Pr(W = w) = 2^{-nR}, \quad w = 1, 2, \ldots, 2^{nR}
  \]
- The $w$th codeword $X^n(w)$ corresponding to the $w$th row of $C$ is sent over the channel.
- The receiver receives a sequence $Y^n$ according to the distribution
  \[
  P(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))
  \]

- We consider jointly typical decoding.
We consider jointly typical decoding.

The receiver declares that the index $\tilde{W}$ was sent if the following conditions are satisfied: $(X^n(\tilde{W}), Y^n)$ is jointly typical and there is no other index $k$, such that $(X^n(k), Y^n) \in A^{(n)}_{c}$.

If no such $\tilde{W}$ exists or if there is more than one such, an error is declared. If $\tilde{W} \neq W$, there is a decoding error.
We will calculate the average probability of error, averaged over all codewords in the codebook and average over all codebooks.

\[
Pr(E) = \sum_{C} P(C) P_{e}(n)(C)
\]

\[
= \sum_{C} P(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_{w}(C)
\]

\[
= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{C} P(C) \lambda_{w}(C)
\]

Due to symmetry of the code construction, the average probability of error averaged over all codes does not depend on particular index that was sent.
Noisy channel coding theorem

We assume without loss of generality that the message $W = 1$ was sent. Average probability of error is given by

$$Pr(E) = \frac{1}{2^{nR}} \sum_{w=1}^{2^n} \sum_{C} P(C) \lambda_w(C)$$

$$= \sum_{C} P(C) \lambda_1(C)$$

$$= Pr(E|W = 1)$$

Let $E_i$ be the event that the $i$th codeword and $Y^n$ are jointly typical.

$$E_i = \left\{ (X^n(i), Y^n) \text{ is in } A_c^{(n)} \right\}, \quad i \in \{1, 2, \ldots, 2^{nR}\}$$
We have $\Pr(E) = \Pr(E|W = 1)$ given by

$$\Pr(E|W = 1) = \Pr(E_1^c \cup E_2 \cup \cdots \cup E_{2^n R} | W = 1) \leq \Pr(E_1^c | W = 1) + \sum_{i=2}^{2^n R} \Pr(E_i | W = 1)$$

By joint AEP, we have $\Pr(E_1^c | W = 1) \leq \epsilon$ for large $n$. 
Noisy channel coding theorem

- We have $Pr(E) = Pr(E|W = 1)$ given by

$$Pr(E|W = 1) = P(E_1^c \cup E_2 \cup \ldots \cup E_{2^nR}|W = 1) \leq P(E_1^c|W = 1) + \sum_{i=2}^{2^nR} P(E_i|W = 1)$$

- By joint AEP, we have $P(E_1^c|W = 1) \leq \epsilon$ for large $n$.
- $X^n(1)$ and $X^n(i)$ are independent, so are $Y^n$ and $X^n(i)$, hence the probability that $Y^n$ and $X^n(i)$ are jointly typical is given by

$$\leq 2^{-n(I(X;Y)-3\epsilon)}.$$
Noisy channel coding theorem

- If $R \leq I(X; Y)$, we can chose $\epsilon$ and $n$ so that the average probability of error is less than $2\epsilon$.

- We choose $p(x)$ in the proof to be the distribution on $X$, $p^*(x)$ that achieves capacity. Then the condition $R < I(X; Y)$ can be replaced by the achievability condition $R < C$. 
Noisy channel coding theorem

- If $R \leq I(X; Y)$, we can chose $\epsilon$ and $n$ so that the average probability of error is less than $2\epsilon$.
- We choose $p(x)$ in the proof to be the distribution on $X$, $p^*(x)$ that achieves capacity. Then the condition $R < I(X; Y)$ can be replaced by the achievability condition $R < C$.
- Since the average probability of error over the codebooks is small, there exists at least one codebook $C^*$ with small average probability of error.
- We throw away the worst half of the codebooks in the best codebook $C^*$. Then we have best half of the codewords having maximal probability of error less than $4\epsilon$. 
Noisy channel coding theorem

- If $R \leq I(X; Y)$, we can chose $\epsilon$ and $n$ so that the average probability of error is less than $2\epsilon$.
- We choose $p(x)$ in the proof to be the distribution on $X$, $p^*(x)$ that achieves capacity. Then the condition $R < I(X; Y)$ can be replaced by the achievability condition $R < C$.
- Since the average probability of error over the codebooks is small, there exists at least one codebook $C^*$ with small average probability of error.
- We throw away the worst half of the codebooks in the best codebook $C^*$. Then we have best half of the codewords having maximal probability of error less than $4\epsilon$.
- If we reindex these codewords, we have $2^{nR-1}$ codewords.

Thus we have constructed a code of rate $R' = R - \frac{1}{n}$ with maximal probability of error $\lambda^{(n)} \leq 4\epsilon$. 
Converse to noisy channel coding theorem

- If information bits from a binary symmetric source (BSS) are sent at a rate \( R \) via a DMC of capacity \( C \) without feedback, then the bit error probability at the destination satisfies

\[
P_b \geq H^{-1}\left(1 - \frac{C}{R}\right), \text{ if } R > C
\]

where \( H^{-1} \) denotes the inverse binary entropy function defined by

\[
H^{-1}(x) = \min\{ p : H(p) = x \}.
\]

Proof:
- Let us consider BSS, i.e. DMS with \( P_U(0) = P_U(1) = 1/2 \).
  Therefore \( H(U) = 1 \) bit.
Converse to noisy channel coding theorem

Proof:
- Let us consider BSS, i.e. DMS with $P_U(0) = P_U(1) = 1/2$. Therefore $H(U) = 1$ bit.
- Also for DMC without feedback,

$$P(y_1, \cdots, y_n|x_1, \cdots, x_N) = \prod_{i=1}^{N} P(y_i|x_i)$$

$$\implies H(Y_1 \cdots Y_N|X_1 \cdots X_N) = \sum_{i=1}^{N} H(Y_i|X_i)$$

Rate of transmission, $R$ is given by $R = K/N$ bits/use.
Converse to noisy channel coding theorem

Proof:

- Let us consider BSS, i.e. DMS with $P_U(0) = P_U(1) = 1/2$. Therefore $H(U) = 1$ bit.
- Also for DMC without feedback,

  
  \[
  P(y_1, \cdots, y_N|x_1, \cdots, x_N) = \prod_{i=1}^{N} P(y_i|x_i)
  \]

  \[
  \implies H(Y_1 \cdots Y_N|X_1 \cdots X_N) = \sum_{i=1}^{N} H(Y_i|X_i)
  \]

- Rate of transmission, $R$ is given by $R = K/N$ bits/use.

- Applying data processing lemma, we get

  \[
  I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \leq I(X_1 \cdots X_K; \hat{Y}_1 \cdots \hat{Y}_K)
  \]

- Applying data processing lemma we also get,

  \[
  I(X_1 \cdots X_N; \hat{Y}_1 \cdots \hat{Y}_K) \leq I(X_1 \cdots X_N; Y_1 \cdots Y_N)
  \]
Converse to noisy channel coding theorem

- Applying data processing lemma we also get,
  \[ I(X_1 \cdots X_N; \hat{U}_1 \cdots \hat{U}_K) \leq I(X_1 \cdots X_N; Y_1 \cdots Y_N) \]

- From above two inequalities, we get
  \[ I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \leq I(X_1 \cdots X_N; Y_1 \cdots Y_N) \]

We can write \( I(X_1 \cdots X_N; Y_1 \cdots Y_N) \) as
\[
I(X_1 \cdots X_N; Y_1 \cdots Y_N) = H(Y_1 \cdots Y_N) - H(Y_1 \cdots Y_N | X_1 \cdots X_N) \\
= H(Y_1 \cdots Y_N) - \sum_{i=1}^{N} H(Y_i | X_i) \\
\leq \sum_{i=1}^{N} [H(Y_i) - H(Y_i | X_i)] \\
= \sum_{i=1}^{N} I(X_i; Y_i) \leq NC
\]
Converse to noisy channel coding theorem

This implies that

\[ I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \leq NC \]

We define bit error probability as

\[ P_b = \frac{1}{K} \sum_{i=1}^{K} P_{ei} \]

where \( P_{ei} = P(\hat{U}_i \neq U_i) \).
Converse to noisy channel coding theorem

- This implies that
  \[ I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \leq NC \]
- We define bit error probability as
  \[ P_b = \frac{1}{K} \sum_{i=1}^{K} P_{ei} \]
  where \( P_{ei} = P(\hat{U}_i \neq U_i) \).
- We can \( H(U_1 \cdots U_K | \hat{U}_1 \cdots \hat{U}_K) \) as
  \[
  H(U_1 \cdots U_K | \hat{U}_1 \cdots \hat{U}_K) = H(U_1 \cdots U_K) - I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \\
  = K - I(U_1 \cdots U_K; \hat{U}_1 \cdots \hat{U}_K) \\
  \geq K - NC \\
  = N(R - C)
  \]

Converse to noisy channel coding theorem

- Also,
  \[
  H(U_1 \cdots U_K | \hat{U}_1 \cdots \hat{U}_K) = \sum_{i=1}^{K} H(U_i | \hat{U}_1 \cdots \hat{U}_K U_1 \cdots U_{i-1}) \\
  \leq \sum_{i=1}^{K} H(U_i | \hat{U}_i)
  \]
Also,

\[ H(U_1 \cdots U_K | \hat{U}_1 \cdots \hat{U}_K) = \sum_{i=1}^{K} H(U_i | \hat{U}_1 \cdots \hat{U}_K U_1 \cdots U_{i-1}) \]

\[ \leq \sum_{i=1}^{K} H(U_i | \hat{U}_i) \]

Hence we get

\[ \sum_{i=1}^{K} H(U_i | \hat{U}_i) \geq N(R - C) \]

Using Fano’s lemma we get

\[ \sum_{i=1}^{K} H(U_i | \hat{U}_i) \leq \sum_{i=1}^{K} H(P_{ei}) \]
Converse to noisy channel coding theorem

• Combining the previous results we get

\[ \frac{1}{K} \sum_{i=1}^{K} H(P_{ei}) \geq \frac{N}{K}(R - C) = 1 - \frac{C}{R} \]

Since \( H(p) \) is a concave function, we get

\[ \frac{1}{K} \sum_{i=1}^{K} H(P_{ei}) \leq H \left( \frac{1}{K} \sum_{i=1}^{K} P_{ei} \right) = H(P_b) \]
Converse to noisy channel coding theorem

1. Combining the previous results we get

\[
\frac{1}{K} \sum_{i=1}^{K} H(P_{ei}) \geq \frac{N}{K} (R - C) = 1 - \frac{C}{R}
\]

2. Since \( H(p) \) is a concave function, we get

\[
\frac{1}{K} \sum_{i=1}^{K} H(P_{ei}) \leq H \left( \frac{1}{K} \sum_{i=1}^{K} P_{ei} \right) = H(P_B)
\]

3. Combining the above results we get

\[
H(P_B) \geq 1 - \frac{C}{R}
\]