4F5: Advanced Wireless Communications
Handout 2: Review of Channel Capacity

Jossy Sayir

Signal Processing and Communications Lab
Department of Engineering
University of Cambridge
jossy.sayir@eng.cam.ac.uk

Lent 2012
Reminder: Basic Block Diagram

...with some more detail (digital communications)
Information Theoretic Channel

Definition (Kelly)

A *channel* is that part of the communication system that one is either unwilling or unable to change.
Outline

1 Definitions and Properties
2 Channel Coding
3 Converse Proof: $R > C, P_e \rightarrow 0$
4 Achievability Proof: $R < C, P_e \rightarrow 0$
5 Summary
Definitions

Entropy, Divergence

Let the random variables $X$, $Y$ take value in the sets $\mathcal{X}$, $\mathcal{Y}$. We define (in bits)

- **Entropy / Uncertainty**

$$H(X) = H(P_X) \overset{\text{def}}{=} - \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) = - \mathbb{E}[\log_2 P_X(X)]$$

- **Divergence / Relative Entropy / Kullback-Leibler “Distance”**

$$D(P_X \| Q_X) \overset{\text{def}}{=} \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{Q_X(x)} = \mathbb{E} \left[ \log_2 \frac{P_X(X)}{Q_X(X)} \right]$$
**Definitions**

**Entropy, Divergence**

Let the random variables $X, Y$ take value in the sets $\mathcal{X}, \mathcal{Y}$. We define (in bits)

- **Entropy / Uncertainty**
  \[
  H(X) = H(P_X) \overset{\text{def}}{=} - \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) = -\mathbb{E}[\log_2 P_X(X)]
  \]

- **Divergence / Relative Entropy / Kullback-Leibler “Distance”**
  \[
  D(P_X \parallel Q_X) \overset{\text{def}}{=} \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{Q_X(x)} = \mathbb{E} \left[ \log_2 \frac{P_X(X)}{Q_X(X)} \right]
  \]
Joint Entropy, Conditional Entropy and Mutual Information

Let the random variables $X$, $Y$ take value in the sets $\mathcal{X}$, $\mathcal{Y}$. We define (in bits)

- **Joint Entropy**
  \[ H(X, Y) \overset{\text{def}}{=} H(P_{XY}) = -\mathbb{E} \left[ \log_2 P_{X,Y}(X,Y) \right] \]

- **Conditional Entropy (conditioned on an event)**
  \[ H(X|Y = y) \overset{\text{def}}{=} H(P_{X|Y=y}) = -\mathbb{E} \left[ \log_2 P_{X|Y}(X|y) \right] \]

- **Conditional Entropy/Equivocation**
  \[ H(X|Y) \overset{\text{def}}{=} \sum_y P_Y(y) H(X|Y = y) = -\mathbb{E} \left[ \log_2 P_{X|Y}(X|Y) \right] \]

- **Mutual Information**
  \[ I(X; Y) \overset{\text{def}}{=} H(X) - H(X|Y) = H(Y) - H(Y|X) \]
  \[ = D(P_{XY}||P_XP_Y) = \mathbb{E} \left[ \log_2 \frac{P_{X,Y}(X,Y)}{P_X(X)P_Y(Y)} \right] \]
Joint Entropy, Conditional Entropy and Mutual Information

Let the random variables $X, Y$ take value in the sets $\mathcal{X}, \mathcal{Y}$. We define (in bits)

- **Joint Entropy**
  
  $$H(X, Y) \overset{\text{def}}{=} H(P_{XY}) = -\mathbb{E}[\log_2 P_{X,Y}(X, Y)]$$

- **Conditional Entropy (conditioned on an event)**
  
  $$H(X|Y = y) \overset{\text{def}}{=} H(P_{X|Y=y}) = -\mathbb{E} [\log_2 P_{X|Y}(X|y)]$$

- **Conditional Entropy/Equivocation**
  
  $$H(X|Y) \overset{\text{def}}{=} \sum_y P_Y(y)H(X|Y = y) = -\mathbb{E}[\log_2 P_{X|Y}(X|Y)]$$

- **Mutual Information**
  
  $$I(X; Y) \overset{\text{def}}{=} H(X) - H(X|Y) = H(Y) - H(Y|X)$$
  
  $$= D(P_{XY}||P_XP_Y) = \mathbb{E} \left[ \log_2 \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)} \right]$$
**Definitions**

### Joint Entropy, Conditional Entropy and Mutual Information

Let the random variables $X$, $Y$ take value in the sets $\mathcal{X}$, $\mathcal{Y}$. We define (in bits)

- **Joint Entropy**

  $$H(X, Y) \overset{\text{def}}{=} H(P_{XY}) = -\mathbb{E}[\log_2 P_{X,Y}(X, Y)]$$

- **Conditional Entropy (conditioned on an event)**

  $$H(X \mid Y = y) \overset{\text{def}}{=} H(P_{X \mid Y = y}) = -\mathbb{E} \left[ \log_2 P_{X \mid Y}(X \mid y) \right]$$

- **Conditional Entropy/Equivocation**

  $$H(X \mid Y) \overset{\text{def}}{=} \sum_y P_Y(y) H(X \mid Y = y) = -\mathbb{E} \left[ \log_2 P_{X \mid Y}(X \mid Y) \right]$$

- **Mutual Information**

  $$I(X; Y) \overset{\text{def}}{=} H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

  $$= D(P_{XY} \parallel P_X P_Y) = \mathbb{E} \left[ \log_2 \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)} \right]$$
Definitions

Joint Entropy, Conditional Entropy and Mutual Information

Let the random variables $X$, $Y$ take value in the sets $\mathcal{X}$, $\mathcal{Y}$. We define (in bits)

- **Joint Entropy**

$$H(X, Y) \overset{\text{def}}{=} H(P_{XY}) = -\mathbb{E}[\log_2 P_{X,Y}(X, Y)]$$

- **Conditional Entropy (conditioned on an event)**

$$H(X|Y = y) \overset{\text{def}}{=} H(P_{X|Y=y}) = -\mathbb{E} [\log_2 P_{X|Y}(X|y)]$$

- **Conditional Entropy/Equivocation**

$$H(X|Y) \overset{\text{def}}{=} \sum_y P_Y(y) H(X|Y = y) = -\mathbb{E}[\log_2 P_{X|Y}(X|Y)]$$

- **Mutual Information**

$$I(X; Y) \overset{\text{def}}{=} H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$= D(P_{XY} \parallel P_X P_Y) = \mathbb{E} \left[ \log_2 \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)} \right]$$
Properties of Entropy, Mutual Information and Relative Entropy

1. Chain rules

\[ H(X, Y) = H(X) + H(Y | X) \]
\[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1) \]

where \( I(X_2; Y | X_1) \) is defined as \( H(X_2 | X_1) - H(X_2 | X_1, Y) \).

2. Positiveness

**entropy**: \( H(X) \geq 0 \), with equality iff \( X \) is deterministic.

implies positiveness of conditional entropy.

**relative entropy**: \( D(P_X || Q_X) \geq 0 \), with equality iff \( P_X = Q_X \)

implies \( I(X; Y) \geq 0 \) (equality iff \( X \) and \( Y \) are independent).

3. Conditioning reduces entropy

\[ H(X | Y) \leq H(X) \]

4. Maximum entropy

\[ H(X) \leq \log |\mathcal{X}|, \quad \text{with equality iff } X \text{ is uniform} \]
Properties of Entropy, Mutual Information and Relative Entropy

1. Chain rules

\[ H(X, Y) = H(X) + H(Y|X) \]
\[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \]

where \( I(X_2; Y|X_1) \overset{\text{def}}{=} H(X_2|X_1) - H(X_2|X_1 Y) \).

2. Positiveness

**Entropy:** \( H(X) \geq 0 \), with equality iff \( X \) is deterministic

implies positiveness of conditional entropy.

**Relative entropy:** \( D(P_X||Q_X) \geq 0 \), with equality iff \( P_X = Q_X \)

implies \( I(X; Y) \geq 0 \) (equality iff \( X \) and \( Y \) are independent).

3. Conditioning reduces entropy

\( H(X|Y) \leq H(X) \)

4. Maximum entropy

\( H(X) \leq \log |\mathcal{X}| \), with equality iff \( X \) is uniform
Properties of Entropy, Mutual Information and Relative Entropy

1. Chain rules

\[ H(X, Y) = H(X) + H(Y|X) \]
\[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \]

where \( I(X_2; Y|X_1) \) def = \( H(X_2|X_1) - H(X_2|X_1 Y) \).

2. Positiveness

entropy: \( H(X) \geq 0 \), with equality iff \( X \) is deterministic

implies positiveness of conditional entropy.

relative entropy: \( D(P_X||Q_X) \geq 0 \), with equality iff \( P_X = Q_X \)

implies \( I(X; Y) \geq 0 \) (equality iff \( X \) and \( Y \) are independent).

3. Conditioning reduces entropy

\[ H(X|Y) \leq H(X) \]

4. Maximum entropy

\[ H(X) \leq \log |\mathcal{X}|, \quad \text{with equality iff } X \text{ is uniform} \]
Properties

Properties of Entropy, Mutual Information and Relative Entropy

1. Chain rules

\[ H(X, Y) = H(X) + H(Y|X) \]
\[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \]

where \( I(X_2; Y|X_1) \) is defined as \( H(X_2|X_1) - H(X_2|X_1 Y) \).

2. Positiveness

**Entropy:** \( H(X) \geq 0 \), with equality iff \( X \) is deterministic

Implies positiveness of conditional entropy.

**Relative entropy:** \( D(P_X||Q_X) \geq 0 \), with equality iff \( P_X = Q_X \)

Implies \( I(X; Y) \geq 0 \) (equality iff \( X \) and \( Y \) are independent).

3. Conditioning reduces entropy

\[ H(X|Y) \leq H(X) \]

4. Maximum entropy

\[ H(X) \leq \log |\mathcal{X}|, \quad \text{with equality iff } X \text{ is uniform} \]
Data Processing Inequality

Let $X \rightarrow Y \rightarrow Z$ form a Markov chain (i.e., $P_{X|Y} = P_{X|Y}P_{Z|Y}$). Then

$$I(X; Z) \leq I(X; Y)$$
Channel Coding

Channel Definitions

- Channel input $X$ over alphabet $\mathcal{X}$.
- Channel output $Y$ over alphabet $\mathcal{Y}$.
- Sequence of transition probabilities

$$\{ P_{Y|X}(y_1, \ldots, y_n|x_1, \ldots, x_n) : n = 1, 2, \ldots \}$$

- Memoryless channel: for $x \in \mathcal{X}^n$, $y \in \mathcal{Y}^n$

$$P_{Y|X}(y|x) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$$

- Discrete Memoryless Channel ($|\mathcal{X}|, |\mathcal{Y}| < \infty$) defined by transition matrix $P$

$$[P]_{i,j} = \Pr(Y = i|X = j)$$

Diagram:

- Binary Symmetric Channel (BSC)
- Binary Erasure Channel (BEC)
A channel coding scheme, or \textit{block code}, is defined by

- A codebook $C \subseteq \mathcal{X}^n$;
- a uniformly distributed message $m \in \mathcal{M} = \{1, \ldots, |\mathcal{M}|\}$ (note that $|C| \leq |\mathcal{M}|$);
- the sequences $x \in C$ are called codewords;
- the coding rate $R = \frac{1}{n} \log_2 |\mathcal{M}|$ (bits/channel use);
- an encoding function $\phi : \mathcal{M} \rightarrow C$ such that $\phi(m) = x_m$ is the codeword corresponding to message $m \in \mathcal{M}$;
- a decoding function $\varphi : \mathcal{Y}^n \rightarrow \mathcal{M}$ such that $\varphi(y) = \hat{m}$ maps the received sequence to an estimated information message.

\[ P_{Y|X}(y|x_m) \]
Channel Coding

Example

Consider a binary ($\mathcal{X} = \{0, 1\}$) code $C$ of length $n = 4$ defined as

$$C = \{(0, 0, 0, 0), (0, 0, 1, 1), (1, 1, 0, 0), (1, 1, 1, 1)\}$$

The rate of the code is $R = \frac{1}{4} \log_2 |\mathcal{M}| = \frac{1}{4} \log_2 4 = \frac{1}{2}$. The message set $\mathcal{M} = \{1, 2, 3, 4\}$ can be represented by 2 bits. Hence the encoder has as input 2 bits and outputs 4 bits (adds redundancy).

Error Probability

The average message (or codeword) error probability of the code $C$ is defined as

$$P_e \triangleq \frac{1}{|\mathcal{M}|} \sum_{m=1}^{\mathcal{M}} P_e(m) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{\mathcal{M}} \sum_{y: \phi(y) \neq m} P_{Y|X}(y|x_m = \phi(m))$$
Consider a binary ($\mathcal{X} = \{0, 1\}$) code $C$ of length $n = 4$ defined as

$C = \{(0, 0, 0, 0), (0, 0, 1, 1), (1, 1, 0, 0), (1, 1, 1, 1)\}$

The rate of the code is $R = \frac{1}{4} \log_2 |\mathcal{M}| = \frac{1}{4} \log_2 4 = \frac{1}{2}$. The message set $\mathcal{M} = \{1, 2, 3, 4\}$ can be represented by 2 bits. Hence the encoder has as input 2 bits and outputs 4 bits (adds redundancy).

The average message (or codeword) error probability of the code $C$ is defined as

$$P_e \triangleq \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} P_e(m) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} \sum_{y: \phi(y) \neq m} P_{Y|X}(y|x_m = \phi(m))$$
Achievable Rates and Capacity

- A rate $R$ is said to be achievable if there exist codes $C$ of length $n$ equipped with encoding and decoding functions $\phi, \varphi$ such that, for every $\epsilon > 0$ and every $n \geq n_\epsilon$ (for some $n_\epsilon$),

\[
\frac{1}{n} \log_2 |\mathcal{M}| \geq R \quad \text{and} \quad P_e \leq \epsilon
\]

- The channel capacity $C$ is defined as the supremum of all achievable rates.

- Thus, for transmission rates $R < C$ there exist coding schemes with arbitrarily small error probability (for sufficiently large block length), while for $R > C$ there exist no such schemes.

Theorem (Shannon’s noisy channel coding theorem)

The channel capacity for a memoryless channel $P_{Y|X}(\cdot)$ is given by

\[
C = \max_{P_X(\cdot)} I(X; Y)
\]

where the maximisation is over all probability distributions on the channel input $X$. 
Channel Capacity

Example

- **BSC**
  \[ C = 1 - H_b(p), \quad P^*_X(0) = P^*_X(1) = \frac{1}{2} \]

- **BEC**
  \[ C = 1 - p, \quad P^*_X(0) = P^*_X(1) = \frac{1}{2} \]
Channel Capacity

Example (AWGN Channel)

- AWGN channel with noise power $\sigma^2$ and input power constraint $P$, i.e.,

  $$P_{Y|X}(y|x) = \frac{1}{\pi\sigma^2} e^{-\frac{|y-x|^2}{\sigma^2}}, \quad x, y \in \mathbb{C} \quad \text{and} \quad \mathbb{E}[|X|^2] \leq P$$

- Capacity is

  $$C = \log_2 (1 + \text{SNR}), \quad \text{SNR} = \frac{P}{\sigma^2} \quad \quad P_X^*(x) = \frac{1}{\pi P} e^{-\frac{|x|^2}{P}}$$

- Gaussian inputs are not practical; we commonly resort to modulations such as PSK/QAM, assuming $P_X(x) = \frac{1}{|\mathcal{X}|}$, $x \in \mathcal{X}$:

  $$I(X; Y) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|x)}{\frac{1}{|\mathcal{X}|} \sum_{x' \in \mathcal{X}} P_{Y|X}(Y|x')} \right]$$
Channel Capacity

AWGN Channel

![Graph showing channel capacity vs. SNR for different modulation schemes]
Simulate the mutual information curves for BPSK, QPSK, 8-PSK and 16-QAM.

Let $P_{Y|X}(y|x) = \frac{1}{\pi\sigma^2} e^{-\frac{1}{\sigma^2} |y-x|^2}$, $x, y \in \mathbb{C}$.

For each SNR value

- calculate the expectation over $X$ as $\frac{1}{M} \sum_{x \in \mathcal{X}} x$
- calculate the expectation over $Y|X$ by randomly generating noise samples $\sim \mathcal{N}_\mathbb{C}(0, \sigma^2)$ and average Montecarlo

end for

TIP: Normalise your constellations in energy, i.e., $\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} |x|^2 = 1$
Converse: \( R > C, P_e \nrightarrow 0 \)

Converse part of Shannon’s Noisy Coding Theorem

- We will show that rates \( R > C \) are not achievable, i.e., if \( R > C \) then \( P_e \) does not tend to zero as \( n \rightarrow \infty \).
- In other words, in order to have \( P_e \rightarrow 0 \) we must have that \( R \leq C \).
- The proof is based on Fano’s inequality and the Data Processing Inequality.
- *Fano’s inequality* relates probability of error and equivocation. We will derive it in two different manners over the next few slides.
Fano’s Inequality

Error Probability and Equivocation

- Suppose we guess $X$ from the observation $Y$. Let the guess be $\hat{X} = g(Y)$.

- **Conditional probability of error (conditioned on $Y = y$):**
  \[
  P_e(y) = \Pr(X \neq \hat{X} | Y = y) = \sum_{x \in \mathcal{X}: x \neq g(y)} P_{X|Y}(x|y)
  \]

- **Probability of error:**
  \[
  P_e = P(X \neq \hat{X}) = \sum_{y \in \mathcal{Y}} P_Y(y) P_e(y)
  \]

- **Equivocation:**
  \[
  H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y)
  \]
Fano’s Inequality

Maximising $H(X \mid Y = y)$ for a given $P_e(y)$

\[
P_e(y) = \sum_{\substack{x \in \mathcal{X} : \\ x \neq g(y)}} P_{X \mid Y}(x \mid y) = 1 - P_{X \mid Y}(g(y) \mid y) \geq 1 - \max_{x \in \mathcal{X}} P_{X \mid Y}(x \mid y)
\]

Thus, for each $y \in \mathcal{Y}$, the probability of error $P_e(y)$ is minimised for

\[
\hat{x} = g(y) = \arg \max_{x \in \mathcal{X}} P_{X \mid Y}(x \mid y),
\]

for which $\max_{x \in \mathcal{X}} P_{X \mid Y}(x \mid y) = 1 - P_e$.

- Maximising $H(X \mid Y = y)$ for a given $P_e(y)$ is equivalent to maximising over all distributions for which $\max_{x \in \mathcal{X}} P_{X \mid Y}(x \mid y) = 1 - P_e$.
- Entropy is maximised by the distribution that is uniform over the remaining symbols in the alphabet $\mathcal{X}$.
Fano’s Inequality

Maximising $H(X\mid Y = y)$ for a given $P_e(y)$

$$P_e(y) = \sum_{\substack{x \in \mathcal{X} : \\ x \neq g(y)}} P_x(y \mid x)$$

$$= 1 - P_x(y \mid g(y))$$

$$\geq 1 - \max_{x \in \mathcal{X}} P_x(y \mid x)$$

Thus, for each $y \in \mathcal{Y}$, the probability of error $P_e(y)$ is minimised for

$$\hat{x} = g(y) = \arg \max_{x \in \mathcal{X}} P_x(y \mid x),$$

for which $\max_{x \in \mathcal{X}} P_x(y \mid x) = 1 - P_e$.

- Maximising $H(X\mid Y = y)$ for a given $P_e(y)$ is equivalent to maximising over all distributions for which $\max_{x \in \mathcal{X}} P_x(y \mid x) = 1 - P_e$.
- Entropy is maximised by the distribution that is uniform over the remaining symbols in the alphabet $\mathcal{X}'$. 

© Jossy Sayir (CUED)
Fano’s Inequality

Maximising $H(X|Y = y)$ for a given $P_e(y)$

\[
P_e(y) = \sum_{\substack{x \in \mathcal{X}: x \neq g(y)}} P_{X|Y}(x|y)
\]

\[
= 1 - P_{X|Y}(g(y)|y)
\]

\[
\geq 1 - \max_{x \in \mathcal{X}'} P_{X|Y}(x|y)
\]

Thus, for each $y \in \mathcal{Y}$, the probability of error $P_e(y)$ is minimised for

\[
\hat{x} = g(y) = \arg \max_{x \in \mathcal{X}'} P_{X|Y}(x|y),
\]

for which $\max_{x \in \mathcal{X}'} P_{X|Y}(x|y) = 1 - P_e$.

- Maximising $H(X|Y = y)$ for a given $P_e(y)$ is equivalent to maximising over all distributions for which $\max_{x \in \mathcal{X}'} P_{X|Y}(x|y) = 1 - P_e$
- Entropy is maximised by the distribution that is uniform over the remaining symbols in the alphabet $\mathcal{X}'$.
Fano's Inequality

\[ P_{X|Y}(x|y) \]

\[ 1 - P_e(y) \]

\[ \frac{P_e(y)}{|X| - 1} \]
Fano’s Inequality

Upper Bound (conditioned on $Y = y$)

\[
H(X|Y = y) \leq -(1 - P_e(y)) \log_2(1 - P_e(y)) - (|\mathcal{X}| - 1) \frac{P_e(y)}{|\mathcal{X}| - 1} \log_2 \frac{P_e(y)}{|\mathcal{X}| - 1}
\]

\[
= H_b(P_e(y)) + P_e \log_2(|\mathcal{X}| - 1)
\]

(1)

Upper Bound (averaged over $Y$)

\[
H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y)
\]

\[
\leq \sum_{y \in \mathcal{Y}} P_Y(y) \left[ H_b(P_e(y)) + P_e(y) \log_2(|\mathcal{X}| - 1) \right]
\]

\[
\leq H_b \left( \sum_{y \in \mathcal{Y}} P_Y(y) P_e(y) \right) + \log_2(|\mathcal{X}| - 1) \sum_{y \in \mathcal{Y}} P_Y(y) P_e(y)
\]

\[
= H_b(P_e) + P_e \log_2(|\mathcal{X}| - 1)
\]

where the third step follows by the concavity of $H_b(\cdot)$ an by Jensen’s inequality.

⇒ This can be weakened to $H(X|Y) \leq 1 + P_e \log_2 |\mathcal{X}|$. 
Fano’s Inequality

Standard proof of Fano’s Inequality

The following alternative proof is much simpler but gives less insight than the one stated previously:

- Let $E$ be an indicator random variable whose value is 0 if $\hat{X} = X$ and 1 if $\hat{X} \neq X$. Note that $P_E(1) = P_e$ and $H(E) = H_b(P_e)$.
- We have

$$
H(X|Y) = H(X, E|Y) - H(E|X, Y) \quad \text{chain rule for entropies}
$$

$$
= H(X, E|Y) \quad \text{because } Y \text{ and } X \text{ determine } E
$$

$$
= H(E|Y) + H(X|Y, E) \quad \text{chain rule for entropies}
$$

$$
\leq H(E) + P_E(0)H(X|Y, E = 0) + P_E(1)H(X|Y, E = 1)
$$

$$
\leq H_b(P_e) + P_e \log_2(|\mathcal{X}| - 1)
$$

Here the fourth step follows because conditioning reduces entropy, and the last step follows because

1. given $E = 0$, $g(Y) = \hat{X}$ determines $X$, which implies that $H(X|Y, E = 0) = 0$
2. given $Y$ and $E = 1$, $X$ can take on at most $|\mathcal{X}| - 1$ values. Hence, its entropy can be at most $\log_2(|\mathcal{X}| - 1)$. 

© Jossy Sayir (CUED) Advanced Wireless Communications Lent 2012 22 / 32
Fano’s Inequality

\[ P_e \leq H(X|\hat{X}) + \log_2 |\mathcal{X}| - \log_2 (|\mathcal{X}| - 1) \]

Allowable region

\[ H(X|\hat{X}) \leq \log_2 |\mathcal{X}| \]
Converse Proof: $R > C$, $P_e \nrightarrow 0$

Proof of the converse part of Shannon’s Noisy Coding Theorem

- Consider the above communications system. We have the Markov chain $m \rightarrow X \rightarrow Y \rightarrow \hat{m}$, where $X = \phi(m)$ and $\hat{m} = \varphi(Y)$.
- If $m$ is drawn uniformly from the message set $\mathcal{M}$, then we have

$$nR = H(m) = H(m|Y) + I(m; Y) \leq 1 + P_e nR + I(m; Y) \leq 1 + P_e nR + I(X; Y) \leq 1 + P_e nR + \sum_{i=1}^{n} I(X_i; Y_i) \leq 1 + P_e nR + nC$$

where $C = \max_{P_X(\cdot)} I(X; Y)$.

- Rewriting the above equation yields

$$P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

Thus, if $R > C$, then $P_e$ does not tend to zero as $n \to \infty$. 

© Jossy Sayir (CUED) Advanced Wireless Communications Lent 2012 24 / 32
Achievability Proof \( R < C, P_e \rightarrow 0 \)

- We show that rates \( R < C \) are achievable, i.e., if \( R < C \) then \( P_e \rightarrow 0 \) as \( n \rightarrow \infty \).
- Codebook construction: generate codewords at random from a particular distribution. We will consider the case where the entries of each of the \(|\mathcal{M}|\) codewords have been generated i.i.d. from \( P_X(\cdot) \), i.e., \( P_X(x) = \prod_{i=1}^{n} P_X(x_i) \).
Achievability Proof $R < C, P_e \rightarrow 0$

We study maximum likelihood decoding, i.e., the decoder chooses the message that maximises the likelihood of having been transmitted

$$\hat{m} = \arg \max_{\mathcal{C}} P_{Y|X}(y|x_m)$$

We study the average error probability over the ensemble of random codes, denoted by

$$\bar{P}_e = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} \bar{P}_e(m)$$

Given the symmetry (random codewords), $\bar{P}_e = \bar{P}_e(m)$ for any $m \in \mathcal{M}$

Averaged over the random code ensemble, we have that

$$\bar{P}_e(m) = \mathbb{E}[\Pr \{\varphi(Y) \neq m | X_m, Y\}]$$

$$= \sum_{x_m} \sum_{y} P_X(x_m) P_{Y|X}(y|x_m) \Pr \{\varphi(y) \neq m | x_m, y\}$$

where $\Pr \{\varphi(y) \neq m | x_m, y\}$ is the probability that, for a channel output $y$, the decoder $\varphi$ selects a codeword other than the transmitted $x_m$.
Achievability Proof $R < C, P_e \rightarrow 0$

We study maximum likelihood decoding, i.e., the decoder chooses the message that maximises the likelihood of having been transmitted

$$\hat{m} = \arg\max_C P_{Y|X}(y|x_m)$$

We study the average error probability over the ensemble of random codes, denoted by $\bar{P}_e = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{\mathcal{M}} P_e(m)$

Given the symmetry (random codewords), $\bar{P}_e = P_e(m)$ for any $m \in \mathcal{M}$

Averaged over the random code ensemble, we have that

$$\bar{P}_e(m) = \mathbb{E}[\Pr \{\varphi(Y) \neq m|X_m, Y\}]$$

$$= \sum_{x_m} \sum_y P_{X}(x_m) P_{Y|X}(y|x_m) \Pr \{\varphi(y) \neq m|x_m, y\}$$

where $\Pr \{\varphi(y) \neq m|x_m, y\}$ is the probability that, for a channel output $y$, the decoder $\varphi$ selects a codeword other than the transmitted $x_m$
Achievability Proof $R < C, P_e \rightarrow 0$

Achievability Proof

- We study maximum likelihood decoding, i.e., the decoder chooses the message that maximises the likelihood of having been transmitted
  \[ \hat{m} = \arg \max_C P_{Y|X}(y|x_m) \]

- We study the average error probability over the ensemble of random codes, denoted by
  \[ \bar{P}_e = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} P_e(m) \]

- Given the symmetry (random codewords), \( \bar{P}_e = P_e(m) \) for any \( m \in \mathcal{M} \)

- Averaged over the random code ensemble, we have that
  \[ \bar{P}_e(m) = \mathbb{E}[\Pr \{ \varphi(Y) \neq m | X_m, Y \}] \]
  \[ = \sum_{x_m} \sum_y P_X(x_m) P_{Y|X}(y|x_m) \Pr \{ \varphi(y) \neq m | x_m, y \} \]

where \( \Pr \{ \varphi(y) \neq m | x_m, y \} \) is the probability that, for a channel output \( y \), the decoder \( \varphi \) selects a codeword other than the transmitted \( x_m \).
Achievability Proof $R < C, P_e \to 0$

**Achievability Proof**

- We study maximum likelihood decoding, i.e., the decoder chooses the message that maximises the likelihood of having been transmitted

  $$\hat{m} = \arg \max_C P_Y|X(y|x_m)$$

- We study the average error probability over the ensemble of random codes, denoted by $\bar{P}_e = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{\mathcal{M}} \bar{P}_e(m)$

- Given the symmetry (random codewords), $\bar{P}_e = \bar{P}_e(m)$ for any $m \in \mathcal{M}$

- Averaged over the random code ensemble, we have that

  $$\bar{P}_e(m) = \mathbb{E}[\Pr \{\varphi(Y) \neq m|X_m, Y\}]$$

  $$= \sum_{x_m} \sum_y P_X(x_m) P_Y|X(y|x_m) \Pr \{\varphi(y) \neq m|x_m, y\}$$

where $\Pr \{\varphi(y) \neq m|x_m, y\}$ is the probability that, for a channel output $y$, the decoder $\varphi$ selects a codeword other than the transmitted $x_m$. 

© Jossy Sayir (CUED)  
Advanced Wireless Communications  
Lent 2012 26 / 32
Achievability Proof $R < C, P_e \to 0$

**Achievability Proof**

- Using the union bound over all possible codewords, and for all $0 \leq \rho \leq 1$,

\[
\Pr \{ \varphi(y) \neq m | x_m, y \} \leq \Pr \left\{ \bigcup_{m' \neq m} \{ \varphi(y) = m' | x_m, y \} \right\} \\
\leq \left( \sum_{m' \neq m} \Pr \{ \varphi(y) = m' | x_m, y \} \right)^\rho
\]

- The pairwise error probability $\Pr \{ \varphi(y) = m' | x_m, y \}$ of wrongly selecting message $m'$ when message $m$ has been transmitted and sequence $y$ has been received is

\[
\Pr \{ \varphi(y) = m' | x_m, y \} = \sum_{x_{m'} : P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m)} P_X(x_{m'})
\]

- Since $P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m)$ and the sum over all $x_{m'}$ upper bounds the sum over the set $\{ x_{m'} : P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m) \}$, for any $s > 0$, the above pairwise error probability can bounded by

\[
\Pr \{ \varphi(y) = m' | x_m, y \} \leq \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(y|x_{m'})}{P_{Y|X}(y|x_m)} \right)^s
\]
Achievability Proof \( R < C, P_e \rightarrow 0 \)

**Achievability Proof**

- Using the union bound over all possible codewords, and for all \( 0 \leq \rho \leq 1 \),

\[
\Pr \{ \varphi(y) \neq m|x_m, y \} \leq \Pr \left\{ \bigcup_{m' \neq m} \{ \varphi(y) = m'|x_m, y \} \right\} \\
\leq \left( \sum_{m' \neq m} \Pr \{ \varphi(y) = m'|x_m, y \} \right)^\rho
\]

- The pairwise error probability \( \Pr \{ \varphi(y) = m'|x_m, y \} \) of wrongly selecting message \( m' \) when message \( m \) has been transmitted and sequence \( y \) has been received is

\[
\Pr \{ \varphi(y) = m'|x_m, y \} = \sum_{x_m': P_{Y|X}(y|x_m') \geq P_{Y|X}(y|x_m)} P_X(x_m')
\]

- Since \( P_{Y|X}(y|x_m') \geq P_{Y|X}(y|x_m) \) and the sum over all \( x_m' \) upper bounds the sum over the set \( \{ x_m' : P_{Y|X}(y|x_m') \geq P_{Y|X}(y|x_m) \} \), for any \( s > 0 \), the above pairwise error probability can bounded by

\[
\Pr \{ \varphi(y) = m'|x_m, y \} \leq \sum_{x_m'} P_X(x_m') \left( \frac{P_{Y|X}(y|x_m')}{P_{Y|X}(y|x_m)} \right)^s
\]
Achievability Proof $R < C$, $P_e \to 0$

**Achievability Proof**

- Using the union bound over all possible codewords, and for all $0 \leq \rho \leq 1$,

\[
\Pr \{ \varphi(y) \neq m | x_m, y \} \leq \Pr \left\{ \bigcup_{m' \neq m} \{ \varphi(y) = m' | x_m, y \} \right\}
\leq \left( \sum_{m' \neq m} \Pr \{ \varphi(y) = m' | x_m, y \} \right)^\rho
\]

- The pairwise error probability $\Pr\{\varphi(y) = m' | x_m, y\}$ of wrongly selecting message $m'$ when message $m$ has been transmitted and sequence $y$ has been received is

\[
\Pr\{\varphi(y) = m' | x_m, y\} = \sum_{x_{m'} : P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m)} P_X(x_{m'})
\]

- Since $P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m)$ and the sum over all $x_{m'}$ upper bounds the sum over the set \{ $x_{m'} : P_{Y|X}(y|x_{m'}) \geq P_{Y|X}(y|x_m)$ \}, for any $s > 0$, the above pairwise error probability can bounded by

\[
\Pr\{\varphi(y) = m' | x_m, y\} \leq \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(y|x_{m'})}{P_{Y|X}(y|x_m)} \right)^s
\]
Achievability Proof $R < C, P_e \to 0$

Achievability Proof

- As $m'$ is a dummy variable, for any $s > 0$ and $0 \leq \rho \leq 1$ it holds that
  
  $$\Pr \{ \varphi(y) \neq m | x_m, y \} \leq \left( (|M| - 1) \sum_{x_m'} P_X(x_m') \left( \frac{P_{Y|X}(y|x_m')}{P_{Y|X}(y|x_m)} \right)^s \right)^\rho$$

- Therefore, we obtain that
  $$\bar{P}_e \leq (|M| - 1)^\rho \mathbb{E} \left[ \left( \sum_{x_m'} P_X(x_m') \left( \frac{P_{Y|X}(y|x_m')}{P_{Y|X}(y|x_m)} \right)^s \right)^\rho \right]$$

- It can be shown (see Gallager 1968) that $s = \frac{1}{1+\rho}$ actually minimises the bound.

- Now, for memoryless channels and input distributions $P_X(x) = \prod_{i=1}^n P_X(x_i)$ we obtain a single-letter characterisation
  $$\bar{P}_e \leq (|M| - 1)^\rho \left( \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(y|x')}{P_{Y|X}(y|x)} \right)^{\frac{1}{1+\rho}} \right)^{\frac{1}{1+\rho}} \right] \right)^n$$
Achievability Proof $R < C, P_e \to 0$

As $m'$ is a dummy variable, for any $s > 0$ and $0 \leq \rho \leq 1$ it holds that

$$\Pr \{ \varphi(y) \neq m|x_m, y \} \leq \left( (|M| - 1) \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(y|x_{m'})}{P_{Y|X}(y|x_m)} \right)^s \right) ^\rho$$

Therefore, we obtain that

$$\bar{P}_e \leq (|M| - 1)^\rho \mathbb{E} \left[ \left( \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(Y|x_{m'})}{P_{Y|X}(Y|X)} \right)^s \right) ^\rho \right]$$

It can be shown (see Gallager 1968) that $s = \frac{1}{1+\rho}$ actually minimises the bound.

Now, for memoryless channels and input distributions $P_X(x) = \prod_{i=1}^n P_X(x_i)$ we obtain a single-letter characterisation

$$\bar{P}_e \leq (|M| - 1)^\rho \left( \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^{\frac{1}{\rho}} \right] \right)^n$$
Achievability Proof $R < C, P_e \to 0$

**Achievability Proof**

- As $m'$ is a dummy variable, for any $s > 0$ and $0 \leq \rho \leq 1$ it holds that

$$\Pr \{\varphi(y) \neq m|x_m, y\} \leq \left((|M| - 1) \sum_{x_{m'}} P(x_{m'}) \left(\frac{P(y|x(y|x_{m'}))}{P(y|x(y|x_m))}\right)^s\right)^\rho$$

- Therefore, we obtain that

$$\bar{P}_e \leq (|M| - 1)^\rho \mathbb{E} \left[\left(\sum_{x_{m'}} P(x_{m'}) \left(\frac{P(y|x(y|x_{m'}))}{P(y|x(y|x_m))}\right)^s\right)^\rho\right]$$

- It can be shown (see Gallager 1968) that $s = \frac{1}{1+\rho}$ actually minimises the bound

- Now, for memoryless channels and input distributions $P_x(x) = \prod_{i=1}^n P_x(x_i)$ we obtain a single-letter characterisation

$$\bar{P}_e \leq (|M| - 1)^\rho \left(\mathbb{E} \left[\left(\sum_{x'} P(x') \left(\frac{P(y|x(x'))}{P(y|x(y|x))}\right)^{\frac{1}{1+\rho}}\right)^\rho\right]\right)^n$$
Achievability Proof $R < C, P_e \rightarrow 0$

**Achievability Proof**

- As $m'$ is a dummy variable, for any $s > 0$ and $0 \leq \rho \leq 1$ it holds that
  \[
  \Pr \{ \varphi(y) \neq m | x_m, y \} \leq \left( (|\mathcal{M}| - 1) \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(y|x_{m'})}{P_{Y|X}(y|x_m)} \right)^s \right)^\rho
  \]

- Therefore, we obtain that
  \[
  \bar{P}_e \leq (|\mathcal{M}| - 1)^\rho \mathbb{E} \left[ \left( \sum_{x_{m'}} P_X(x_{m'}) \left( \frac{P_{Y|X}(Y|x_{m'})}{P_{Y|X}(Y|x_m)} \right)^s \right)^\rho \right]
  \]

- It can be shown (see Gallager 1968) that $s = \frac{1}{1+\rho}$ actually minimises the bound

- Now, for memoryless channels and input distributions $P_X(x) = \prod_{i=1}^{n} P_X(x_i)$ we obtain a single-letter characterisation
  \[
  \bar{P}_e \leq (|\mathcal{M}| - 1)^\rho \left( \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^\rho \right] \right)^n
  \]
Achievability Proof $R < C$, $P_e \to 0$

Hence, since $|\mathcal{M}| = 2^{nR}$ for any input distribution $P_X(x)$, and $0 \leq \rho \leq 1$,

$$P_e \leq 2^{-n(E_0(\rho) - \rho R)} \quad \text{with} \quad E_0(\rho) \triangleq -\log_2 \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^{\rho} \right]$$

$E_0(\rho)$ is called the Gallager function. The expectation is carried out according to the joint distribution $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$.

We define the random coding exponent as

$$E_r(R) \triangleq \max_{0 \leq \rho \leq 1} (E_0(\rho) - \rho R).$$

Hence, the tightest error probability bound is obtained as

$$P_e \leq 2^{-nE_r(R)}$$

The average error probability goes to zero for increasing $n$ when

$$E_0(\rho) > \rho R$$
Achievability Proof $R < C, P_e \to 0$

Hence, since $|\mathcal{M}| = 2^{nR}$ for any input distribution $P_X(x)$, and $0 \leq \rho \leq 1$,

$$\bar{P}_e \leq 2^{-n(E_0(\rho) - \rho R)} \quad \text{with} \quad E_0(\rho) \triangleq -\log_2 \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_Y|X(Y|x')}{P_Y|X(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^{\rho} \right]$$

$E_0(\rho)$ is called the Gallager function. The expectation is carried out according to the joint distribution $P_X, Y(x, y) = P_Y|X(y|x)P_X(x)$.

We define the random coding exponent as

$$E_r(R) \triangleq \max_{0 \leq \rho \leq 1} (E_0(\rho) - \rho R).$$

Hence, the tightest error probability bound is obtained as

$$\bar{P}_e \leq 2^{-nE_r(R)}$$

The average error probability goes to zero for increasing $n$ when

$$E_0(\rho) > \rho R$$
Achievability Proof $R < C, P_e \to 0$

Hence, since $|\mathcal{M}| = 2^{nR}$ for any input distribution $P_X(x)$, and $0 \leq \rho \leq 1$,

$$\bar{P}_e \leq 2^{-n(E_0(\rho) - \rho R)}$$

with $E_0(\rho) \triangleq -\log_2 \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(y|x')}{P_{Y|X}(y|x)} \right)^{\frac{1}{1+\rho}} \right)^\rho \right]$

$E_0(\rho)$ is called the Gallager function. The expectation is carried out according to the joint distribution $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$.

We define the random coding exponent as

$$E_r(R) \triangleq \max_{0 \leq \rho \leq 1} (E_0(\rho) - \rho R).$$

Hence, the tightest error probability bound is obtained as

$$\bar{P}_e \leq 2^{-nE_r(R)}$$

The average error probability goes to zero for increasing $n$ when $E_0(\rho) > \rho R$. 
Achievability Proof $R < C$, $P_e \to 0$

Achievability Proof

- Hence, since $|\mathcal{M}| = 2^{nR}$ for any input distribution $P_X(x)$, and $0 \leq \rho \leq 1$,

  \[ \bar{P}_e \leq 2^{-n(E_0(\rho) - \rho R)} \quad \text{with} \quad E_0(\rho) \triangleq -\log_2 \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(Y|X)}{P_{Y|X}(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^\rho \right] \]

- $E_0(\rho)$ is called the Gallager function. The expectation is carried out according to the joint distribution $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$.

- We define the random coding exponent as

  \[ E_r(R) \triangleq \max_{0 \leq \rho \leq 1} (E_0(\rho) - \rho R). \]

- Hence, the tightest error probability bound is obtained as

  \[ \bar{P}_e \leq 2^{-nE_r(R)} \]

- The average error probability goes to zero for increasing $n$ when

  \[ E_0(\rho) > \rho R \]
Achievability Proof $R < C, P_e \rightarrow 0$

Hence, since $|\mathcal{M}| = 2^{nR}$ for any input distribution $P_X(x)$, and $0 \leq \rho \leq 1$,

$$\bar{P}_e \leq 2^{-n(E_0(\rho) - \rho R)} \quad \text{with} \quad E_0(\rho) \triangleq -\log_2 \mathbb{E} \left[ \left( \sum_{x'} P_X(x') \left( \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right)^{\frac{1}{1+\rho}} \right)^\rho \right]$$

$E_0(\rho)$ is called the Gallager function. The expectation is carried out according to the joint distribution $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$.

We define the random coding exponent as

$$E_r(R) \triangleq \max_{0 \leq \rho \leq 1} (E_0(\rho) - \rho R).$$

Hence, the tightest error probability bound is obtained as

$$\bar{P}_e \leq 2^{-nE_r(R)}$$

The average error probability goes to zero for increasing $n$ when

$E_0(\rho) > \rho R$
Achievability Proof $R < C, P_e \to 0$

Using that $E_0(0) = 0$ we see that
\[
\left. \frac{dE_0(\rho)}{d\rho} \right|_{\rho=0} = \lim_{\rho \to 0} \frac{E_0(\rho)}{\rho} = -\mathbb{E} \left[ \log_2 \sum_{x'} P_X(x') \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right] = \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{\sum_{x'} P_X(x') P_{Y|X}(Y|x')} \right] = I(X; Y)
\]

Actually, $0 < \frac{dE_0(\rho)}{d\rho} \leq I(X; Y)$ with equality iff $\rho = 0$. Also, $\frac{d^2E_0(\rho)}{d\rho^2} \leq 0$ for $\rho \geq 0$

Then $E_0(\rho)$ is an increasing function of $\rho \geq 0$ and its maximum slope is at $\rho = 0$, given by $I(X; Y)$

It follows that the function $g(\rho) = E_0(\rho) - \rho R$ has a maximum in $[0, 1]$ if
\[
\frac{dE_0(\rho)}{d\rho} - R = 0
\]

has a solution in $[0, 1]$. Otherwise, the maximum is achieved by $\rho = 1$ for $R < I(X; Y)$ or $\rho = 0$ for $R \geq I(X; Y)$
Achievability Proof $R < C, P_e \to 0$

Achievability Proof

- Using that $E_0(0) = 0$ we see that

$$\left. \frac{dE_0(\rho)}{d\rho} \right|_{\rho=0} = \lim_{\rho \to 0} \frac{E_0(\rho)}{\rho} = -\mathbb{E} \left[ \log_2 \sum_{x'} P_X(x') \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right]$$

$$= \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{\sum_{x'} P_X(x') P_{Y|X}(Y|x')} \right] = \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] = I(X; Y)$$

- Actually, $0 < \frac{dE_0(\rho)}{d\rho} \leq I(X; Y)$ with equality iff $\rho = 0$. Also, $\frac{d^2E_0(\rho)}{d\rho^2} \leq 0$ for $\rho \geq 0$

- Then $E_0(\rho)$ is an increasing function of $\rho \geq 0$ and its maximum slope is at $\rho = 0$, given by $I(X; Y)$

- It follows that the function $g(\rho) = E_0(\rho) - \rho R$ has a maximum in $[0, 1]$ if

$$\frac{dE_0(\rho)}{d\rho} - R = 0$$

has a solution in $[0, 1]$. Otherwise, the maximum is achieved by $\rho = 1$ for $R < I(X; Y)$ or $\rho = 0$ for $R \geq I(X; Y)$
Achievability Proof $R < C, P_e \to 0$

Achievability Proof

Using that $E_0(0) = 0$ we see that

$$\frac{dE_0(\rho)}{d\rho} \bigg|_{\rho=0} = \lim_{\rho \to 0} \frac{E_0(\rho)}{\rho} = -\mathbb{E} \left[ \log_2 \sum_{x'} P_x(x') \frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)} \right]

= \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{\sum_{x'} P_x(x') P_{Y|X}(Y|x')} \right] = \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] = I(X; Y)

Actually, $0 < \frac{dE_0(\rho)}{d\rho} \leq I(X; Y)$ with equality iff $\rho = 0$. Also, $\frac{d^2E_0(\rho)}{d\rho^2} \leq 0$ for $\rho \geq 0$

Then $E_0(\rho)$ is an increasing function of $\rho \geq 0$ and its maximum slope is at $\rho = 0$, given by $I(X; Y)$

It follows that the function $g(\rho) = E_0(\rho) - \rho R$ has a maximum in $[0, 1]$ if

$$\frac{dE_0(\rho)}{d\rho} - R = 0$$

has a solution in $[0, 1]$. Otherwise, the maximum is achieved by $\rho = 1$ for $R < I(X; Y)$ or $\rho = 0$ for $R \geq I(X; Y)$
Achievability Proof \( R < C, P_e \to 0 \)

Using that \( E_0(0) = 0 \) we see that

\[
\frac{dE_0(\rho)}{d\rho} \bigg|_{\rho=0} = \lim_{\rho \to 0} \frac{E_0(\rho)}{\rho} = -\mathbb{E} \left[ \log_2 \sum_{x'} P_x(x') \frac{P_{Y|X}(Y|X')}{P_{Y|X}(Y|X)} \right]
= \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{\sum_{x'} P_x(x') P_{Y|X}(Y|X')} \right] = \mathbb{E} \left[ \log_2 \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] = I(X; Y)
\]

Actually, \( 0 < \frac{dE_0(\rho)}{d\rho} \leq I(X; Y) \) with equality iff \( \rho = 0 \). Also, \( \frac{d^2E_0(\rho)}{d\rho^2} \leq 0 \) for \( \rho \geq 0 \)

Then \( E_0(\rho) \) is an increasing function of \( \rho \geq 0 \) and its maximum slope is at \( \rho = 0 \), given by \( I(X; Y) \)

It follows that the function \( g(\rho) = E_0(\rho) - \rho R \) has a maximum in \([0, 1]\) if

\[
\frac{dE_0(\rho)}{d\rho} - R = 0
\]

has a solution in \([0, 1]\). Otherwise, the maximum is achieved by \( \rho = 1 \) for \( R < I(X; Y) \) or \( \rho = 0 \) for \( R \geq I(X; Y) \)
Achievability Proof $R < C, P_e \to 0$

- Since $\bar{P}_e \leq 2^{-nE_r(R)}$, then there must exist codes for which $P_e \leq 2^{-nE_r(R)}$
- Finally, since so far $P_X(X)$ was fixed, we can maximise over the input distribution to obtain the tightest bound and prove the achievability part of Shannon’s theorem, i.e., rates $R < C$ are achievable.

Gallager function $E_0(\rho)$ and the random coding error exponent $E_r(R)$ for 16-QAM in an AWGN channel with SNR=5 dB
Channel Capacity Theorem

Summary

We have proved that the channel capacity is

\[
C = \max_{p_X(X)} I(X; Y)
\]

Achievability

For every rate \( R < C \) there exists a codebook \( C \) for which the probability of error tends to zero as \( n \to \infty \).

Converse

The probability of error satisfies

\[
P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}
\]

Thus, if \( R > C \) then the probability of error does not tend to zero as \( n \to \infty \).

Summary

The channel capacity is the fundamental limit of information transmission.