Delivering Information Reliably

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

October 26, 2023

The information processing task motivating the study of this lecture:

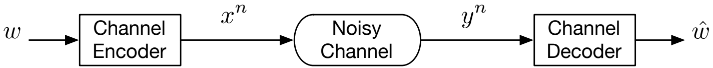
> *For a length-k i.i.d. $\mathrm{Ber}(\frac{1}{2})$ bit sequence, given a noisy channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, design a good encoding scheme to represent the k-bit sequence using n symbols in $\mathcal{X}$, and a good decoding scheme to reconstruct the k-bit sequence reliably from the observed n symbols in $\mathcal{Y}$ at the output of the channel.*

Note: how the encoding and decoding scheme work are known *a priori*.

Fundamental Questions:

- What is the maximum possible ratio $\frac{k}{n}$ (transmission rate) ?
- How to achieve that fundamental limit?

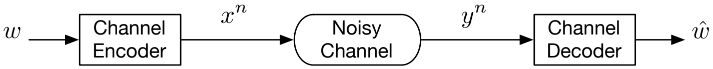
The channel coding problem (Shannon's abstraction)



Meta Description

1. **Message**: Random message $W \sim \mathrm{Unif}\{1, \dots, 2^k\}$.
2. **Channel**: Consist of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and a family of conditional laws $\left\{\mathsf{P}_{Y_i|X^i,Y^{i-1}} \,\middle|\, i \in \mathbb{N}\right\}$ determining the stochastic relationship between the output symbol Y_i and the input symbol X_i along with all past signals (X^{i-1}, Y^{i-1}).
3. **Encoder**: Encode the message w by a length-n codeword $x^n \in \mathcal{X}^n$.
4. **Decoder**: Reconstruct message $\hat{w}$ from the channel output y^n.
5. **Efficiency**: Maximize the code rate $\mathrm{R} := \frac{k}{n}$ bits/channel use, given a certain decoding criterion.

Decoding criterion: small error probability



A key performance measure: Error Probability $\mathsf{P}_{\mathsf{e}}^{(n)} := \Pr\{W \neq \hat{W}\}$.

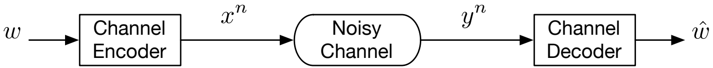
Question: Is it possible to get zero error probability?

Answer: Probably not, unless the channel noise has some special structure.

Following the development of lossless source coding, Shannon turned the attention to answering the following question:

> *Is it possible to have a sequence of encoder/decoder pairs such that $\mathsf{P}_{\mathsf{e}}^{(n)} \to 0$ as $n \to \infty$? If so, what is the largest possible code rate R where vanishing error probability is possible?*

Encoding and decoding



Decoder takes the observation Y^n generated from the noisy channel and make a decision

$$\hat{W} = \mathrm{dec}(Y^n) \in \mathcal{W} \equiv \{1, \ldots, 2^k\}.$$

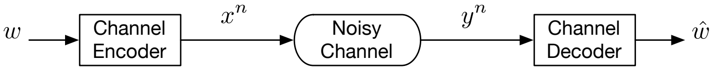
This is just a 2^k-ary detection (hypothesis testing) problem, with

$$\mathcal{H}_w : Y^n \sim \mathsf{P}_w \equiv \mathsf{P}_{Y^n|X^n}(\cdot|x^n(w)),\ w = 1, \ldots, 2^k.$$

Looks pretty simple – just use maximum likelihood.

But the analysis of the error probability is tricky.

Encoding and decoding (cont'd)



The even more tricky part is the encoder.

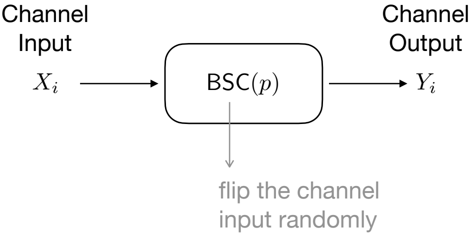
Encoder places the 2^k *codewords* in $\mathcal{X}^n$ so that after passing through the noisy channel $\mathsf{P}_{Y^n|X^n}$, the distributions of the observations

$$\left\{\mathsf{P}_w(\cdot) \equiv \mathsf{P}_{Y^n|X^n}(\cdot|x^n(w)) \,\middle|\, w = 1, \ldots, 2^k\right\}$$

are well-separated.

To make these challenges concrete and see that there are ways to resolve them, let us begin with an example – the binary symmetric channel (BSC), one of the simplest yet non-trivial noisy channel.

Motivating example: the binary symmetric channel


Channel Input

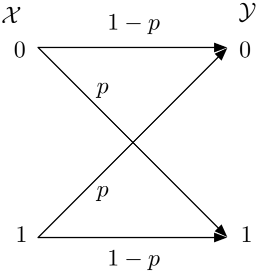
X_i

BSC(p)

Channel Output

Y_i

flip the channel input randomly



- Binary input/output: $x_i, y_i \in \{0,1\}$, $i = 1, 2, \ldots$.
- Channel flips the input bit i.i.d. with probability $p \in (0, 1/2)$:

$$Y_i = X_i \oplus Z_i,\ Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p),\ i = 1, 2, \ldots.$$

Encoding

$$W \equiv \begin{bmatrix} B_1 & B_2 & \dots & B_k \end{bmatrix} \longrightarrow \boxed{\text{enc}} \longrightarrow \begin{bmatrix} X_1 & X_2 & \dots & \dots & X_n \end{bmatrix}$$

message

$$\boldsymbol{B} \sim \mathrm{Unif}\big(\{0,1\}^k\big)$$

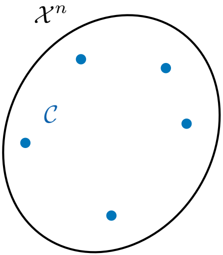
codeword

$$\boldsymbol{X} \in \{0,1\}^n$$

- Goal: introduce redundancy to combat channel noises (for BSC, bit flips).
- Codebook: the range of the encoding function enc:

$$\mathcal{C} = \big\{\mathrm{enc}(w) \equiv \boldsymbol{x}(w) \,\big|\, w \in \mathcal{W} \equiv \{1, \dots, 2^k\}.\big\}.$$

- eg. Repetition coding $\mathcal{C} = \{[0 \ \dots \ 0], [1 \ \dots \ 1]\}$.



Pictorially, the encoding function should place the codewords far apart so that they do not get confused with one another.

Decoding

$$\begin{bmatrix} Y_1 & Y_2 & \dots & \dots & Y_n \end{bmatrix} \longrightarrow \boxed{\text{dec}} \longrightarrow \begin{bmatrix} \hat{B}_1 & \hat{B}_2 & \dots & \hat{B}_k \end{bmatrix} \equiv \hat{W}$$

received signal

$\boldsymbol{Y} \in \{0,1\}^n$

decoded message

$\hat{W} \in \mathcal{W}$

- Performance metic: probability of error $\mathsf{P}_{\mathsf{e}}^{(n)} = \mathsf{Pr}\{\hat{W} \neq W\}$.
- Since $W \sim \mathrm{Unif}(\mathcal{W})$ (uniform prior), *maximum likelihood* decoding is the optimal: $\hat{w}_{\mathrm{ML}}(\boldsymbol{y}) = \arg\max_{w \in \mathcal{W}} \mathsf{P}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}\middle|\boldsymbol{x}(w)\right)$.
- For BSC, likelihood function is simple and determined by the *Hamming distance* between $\boldsymbol{y}$ and the candidate codeword $\boldsymbol{x}(w)$:

$$\mathsf{P}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}\middle|\boldsymbol{x}(w)\right) = p^{\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))}(1-p)^{n-\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))} = (1-p)^n \left(\tfrac{p}{1-p}\right)^{\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))}.$$

Since $0 < p < 1/2$, we have

$$\hat{w}_{\mathrm{ML}}(\boldsymbol{y}) = \arg\min_{w \in \mathcal{W}} \mathrm{d_H}\left(\boldsymbol{y}, \boldsymbol{x}(w)\right)$$

Error probability analysis

Under ML decoding, the error probability is given as

$$\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)} = \frac{1}{2^k} \sum_{w\in\mathcal{W}} \underbrace{\Pr\left\{\exists\, \tilde{w} \neq w \text{ such that } \mathrm{d_H}(\boldsymbol{Y}, \boldsymbol{x}(\tilde{w})) \le \mathrm{d_H}(\boldsymbol{Y}, \boldsymbol{x}(w)) \,\middle|\, W = w\right\}}_{\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(w)}$$

The event that $\boldsymbol{x}(\tilde{w})$ is closer to $\boldsymbol{Y}$ than the actual $\boldsymbol{x}(w)$ is just the event that BSC flips more than half of the bits at which $\boldsymbol{x}(w)$ and $\boldsymbol{x}(\tilde{w})$ differ.

Hence by the union bound, we may get to

$$\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(w) \le \sum_{\substack{\tilde{w}\in\mathcal{W}\\ \tilde{w}\neq w}} \sum_{j=\left\lceil \frac{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}),\boldsymbol{x}(w))}{2} \right\rceil}^{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}),\boldsymbol{x}(w))} \binom{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}),\boldsymbol{x}(w))}{j} p^j (1-p)^{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}),\boldsymbol{x}(w))-j}.$$

It largely depends on how the encoding function populates the codewords in the codebook and hard to analyze, let alone the asymptote as $n \to \infty$.

Reflections

- Recall: for lossless source coding, we leveraged the *concentration of probability* (AEP) when $n \to \infty$ to achieve data compression.

  Reason: distribution of the random source is known and "well-behaved".

- But for the channel coding problem, the encoding function enc makes the distribution of coded symbols non-i.i.d. in general, and it becomes hard to control the statistical behavior of the codewords.

$$\underset{\text{i.i.d. over time}}{\boldsymbol{B}} \longrightarrow \boxed{\mathrm{enc}} \longrightarrow \boldsymbol{x}(\boldsymbol{B}) = \underset{\text{not i.i.d. over time}}{\begin{bmatrix} x_1(\boldsymbol{B}) & x_2(\boldsymbol{B}) & \dots & x_n(\boldsymbol{B}) \end{bmatrix}}$$

Work-around:

1. "Random" encoding
2. "Typicality" decoding

Typicality decoding

Observation: bit-flips are i.i.d. $\mathrm{Ber}(p)$, and hence the number of bit-flips follows $\mathrm{Binom}(n,p)$ and concentrates at np when $n \to \infty$. So, it is "typical" that there are $\approx np$ bit-flips, that is, $\forall\, \varepsilon > 0$,

$$\mathsf{Pr}\Big\{ \overbrace{\mathrm{w}(\boldsymbol{Z})}^{\text{\# of bit-flips}} \le n(p+\varepsilon) \Big\} \to 1 \quad \text{as } n \to \infty.$$

Typicality decoding:

$$\hat{w}_{\mathrm{T}} = \text{the unique } w \in \mathcal{W} \text{ such that } \mathrm{d}_{\mathrm{H}}(\boldsymbol{y}, \boldsymbol{x}(w)) \le n(p+\varepsilon).$$

Obviously, $\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{ML}}(w) \le \mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(w)$. Furthermore, by the union bound,

$$\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(w) \le \underbrace{\mathsf{Pr}\Big\{ \overbrace{\mathrm{w}(\boldsymbol{Z})}^{\mathrm{d}_{\mathrm{H}}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(w))} > n(p+\varepsilon) \Big\}}_{\to 0 \text{ as } n \to \infty}$$

$$+ \textstyle\sum_{\tilde{w} \in \mathcal{W}, \tilde{w} \ne w} \mathsf{Pr}\Big\{ \mathrm{d}_{\mathrm{H}}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(\tilde{w})) \le n(p+\varepsilon) \Big\} \qquad (*)$$

Random encoding

$$\mathsf{Pr}\left\{\mathrm{d_H}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(\tilde{w})) \le n(p+\varepsilon)\right\}$$

This term remains difficult to analyze as it depends on the structure of the codebook (more specifically, pairwise Hamming distance between codewords).

To overcome the difficulty, Shannon came up with a *random coding* idea:

> *To prove the existences of codebook with* $\mathsf{P}_{\mathsf{e}}^{(n)} \le \epsilon$*, it suffices to show that over a set of codebooks, the averaged* $\mathsf{P}_{\mathsf{e}}^{(n)} \le \epsilon$*.*

One may consider the codebook $\mathcal{C} = \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(2^k)\}$ as a $2^k \times n$ matrix

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

Random codebook simply means that the codebook matrix is random and follows a certain distribution: $\mathbf{C} \sim \mathsf{P}_{\mathbf{C}} \in \mathcal{P}(\{0,1\}^{2^k \times n})$.

With the random codebook matrix $\mathbf{C}$, the goal now turns to proving

$$\mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}},W\sim\mathrm{Unif}(\mathcal{W})}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(W;\mathbf{C})\right] \le \epsilon.$$

Clearly, $\mathbf{C}$ is chosen to be independent of W. Hence, it suffices to show that

$$\forall\, w \in \mathcal{W},\ \mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(w;\mathbf{C})\right] \le \epsilon.$$

It boils down to show that $\mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}}}\left[(*)\right] \le \epsilon/2$, that is,

$$\textstyle\sum_{\tilde{w}\in\mathcal{W},\tilde{w}\neq w} \mathsf{P}_{\mathbf{C},\boldsymbol{Z}}\left\{\underbrace{\mathrm{d_H}(\boldsymbol{X}(w)\oplus\boldsymbol{Z},\boldsymbol{X}(\tilde{w}))}_{\mathrm{w}(\boldsymbol{X}(w)\oplus\boldsymbol{Z}\oplus\boldsymbol{X}(\tilde{w}))} \le n(p+\varepsilon)\right\} \le \epsilon/2.$$

Let’s pick a distribution on $\mathbf{C}$ so that the analysis becomes simple:

$$X_i(w) \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2) \quad \forall\, i = 1, 2, \ldots, n,\ \forall\, w = 1, 2, \ldots, 2^k,$$

that is, the entries of $\mathbf{C}$ are i.i.d. $\mathrm{Ber}(1/2)$. Then, each codeword (row of $\mathbf{C}$) is a length-n i.i.d. $\mathrm{Ber}(1/2)$ sequence.

Since $\boldsymbol{X}(w), \boldsymbol{X}(\tilde{w}), \boldsymbol{Z}$ are independent, $\boldsymbol{X}(w) \oplus \boldsymbol{Z} \oplus \boldsymbol{X}(\tilde{w})$ is also a length-n i.i.d. $\mathrm{Ber}(1/2)$ sequence.

As a result, $\forall\, \tilde{w} \neq w$,

$$\begin{aligned} &\mathsf{P}_{\mathbf{C},\boldsymbol{Z}}\left\{\mathrm{w}(\boldsymbol{X}(w) \oplus \boldsymbol{Z} \oplus \boldsymbol{X}(\tilde{w})) \le n(p+\varepsilon)\right\} \\ &= \mathsf{Pr}\left\{\mathrm{Binom}(n, \tfrac{1}{2}) \le n(p+\varepsilon)\right\} \le 2^{-n\mathsf{d}_\mathsf{b}\left(p+\varepsilon \,\middle\|\, \frac{1}{2}\right)} \qquad \text{(Chernoff Bound)} \end{aligned}$$

Wrapping-up: achievability

Putting everything together, we show that $\exists$ a codebook $\mathcal{C}$ of size 2^k such that

$$\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)} \le \epsilon/2 + (2^k - 1)2^{-n\mathsf{d_b}\left(p+\varepsilon \,\middle\|\, \frac{1}{2}\right)}.$$

A sufficient condition for the second term to vanish is $k < n\mathsf{d_b}\left(p + \varepsilon \,\middle\|\, \frac{1}{2}\right)$.

Choosing $\varepsilon > 0$ judiciously close to 0, we can show that $\forall\, \delta > 0$ and $\forall\, \epsilon \in (0,1)$, there exists a codebook $\mathcal{C}$ of size 2^k such that

$$k > n\left(\mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right) - \delta\right) \quad \text{and} \quad \mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)} \le \epsilon.$$

Hence,

$$\liminf_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) \ge \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right),$$

establishing the achievability part of the coding theorem for $\mathsf{BSC}(p)$.

Optimality

Is it true that $\lim_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \frac{1}{2}\right) = \mathrm{D}\left(\mathrm{Ber}(p) \middle\| \mathrm{Ber}(\frac{1}{2})\right)$?

The answer is **yes**. The divergence reminds us of binary hypothesis testing, and the proof below is built on the idea of *testing between two channel laws*.

For a given encoding/decoding pair $(\mathrm{enc}, \mathrm{dec})$ with $\mathsf{P}_\mathsf{e}^{(n)} \le \epsilon$, consider a binary hypothesis testing problem

$$\mathcal{H}_0 : \quad W \to \boxed{\mathrm{enc}} \to \boldsymbol{X} \to \boxed{\mathsf{BSC}(p)^{\otimes n}} \to \boldsymbol{Y} \to \boxed{\mathrm{dec}} \to \hat{W}$$

$$\mathcal{H}_1 : \quad W \to \boxed{\mathrm{enc}} \to \boldsymbol{X} \to \boxed{\mathsf{BSC}(1/2)^{\otimes n}} \to \boldsymbol{Y} \to \boxed{\mathrm{dec}} \to \hat{W}$$

with observation tuple $(W, \boldsymbol{X}, \boldsymbol{Y}, \hat{W})$ following P_0 and P_1 respectively.

Key observation 1: the second channel $\mathsf{BSC}(1/2)$ completely breaks the dependency between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and hence the decoding performance is the same as pure random guess, that is, $\mathsf{P}_1\{W = \hat{W}\} = \frac{1}{2^k}$.

For the above HT problem, one can use a test that accept $\mathcal{H}_0$ iff $W = \hat{W}$, and this test has

$$\pi_{1|0}^{(n)} = \mathsf{P}_0\{W \neq \hat{W}\} \equiv \mathsf{P}_{\mathsf{e}}^{(n)} \leq \epsilon$$
$$\pi_{0|1}^{(n)} = \mathsf{P}_1\{W = \hat{W}\} = \tfrac{1}{2^k}.$$

With the language of binary HT, we immediately have $\frac{1}{2^k} \geq \varpi_{0|1}^*(n,\epsilon)$.

Key observation 2: the above binary HT problem is equivalent to

$$\begin{aligned} \mathcal{H}_0 : &\quad Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p),\ i = 1, 2, \ldots, n \\ \mathcal{H}_1 : &\quad Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\tfrac{1}{2}),\ i = 1, 2, \ldots, n. \end{aligned}$$

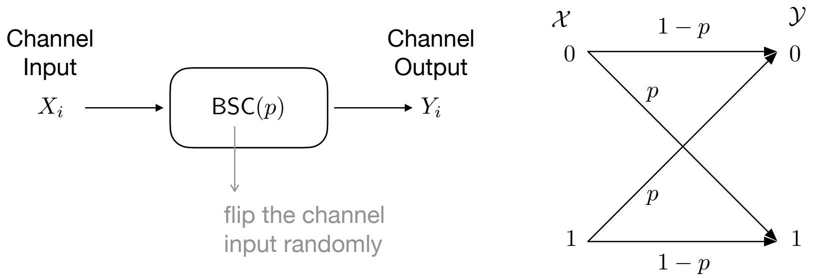
Hence, by the Chernoff-Stein Lemma, $\lim_{n\to\infty} -\frac{1}{n} \log \varpi_{0|1}^*(n,\epsilon) = \mathsf{d_b}\big(p \,\big\|\, \frac{1}{2}\big)$.

Combining everything together, we show that

$$\limsup_{n\to\infty} \tfrac{1}{n} k^*(n,\epsilon) \leq \lim_{n\to\infty} -\tfrac{1}{n} \log \varpi_{0|1}^*(n,\epsilon) = \mathsf{d_b}\big(p \,\big\|\, \tfrac{1}{2}\big),$$

and the optimal rate for communication over the BSC is characterized.

Summary



For the binary symmetric channel with flip probability p,

$$\lim_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right) = 1 - \mathsf{H_b}(p) \quad \forall\, \epsilon \in (0,1).$$

where $k^*(n,\epsilon)$ is the largest k such that there exists channel codes of codebook size 2^k and error probability $\le \epsilon$.

How to extend the result to general channels?

Outline

In this lecture, we will show that (for discrete memoryless channels) the fundamental limit is the channel capacity, a quantity that can be computed by maximizing the "mutual information" between the input and the output of the channel, when we want to reconstruct the bits with vanishing error probability.

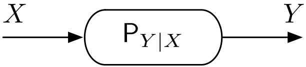
1. We begin with an intuitive motivation to introduce mutual information.
   - Motivation: single use of a channel.
   - Data processing inequality, chain rule, convexity.

2. Next we prove Shannon's noisy channel coding theorem.
   - Use Fano's inequality to prove the converse part.
   - Use typicality arguments to prove the achievability part.
   - Source-channel separation.

1. Mutual Information
   - Definitions
   - Properties

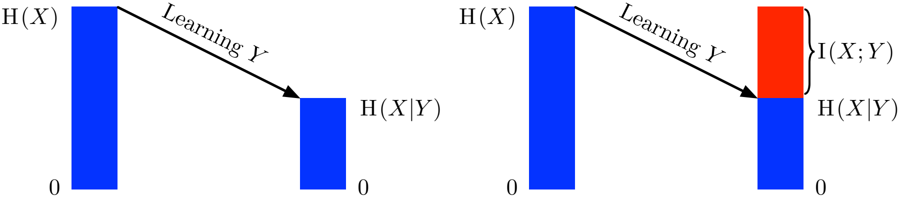
2. Noisy Channel Coding Theorem
   - Channel Capacity
   - Proof of Weak Converse
   - Feedback Capacity
   - Proof of Achievability
   - Source-Channel Separation

Consider a single use of the channel $\mathsf{P}_{Y|X}$:



$\mathrm{H}(X)$ quantifies the amount of uncertainty of the input X.

$\mathrm{H}(X|Y)$ quantifies the amount of uncertainty of X given the output Y.



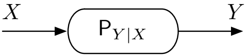
Question: How much information does Y tell about X through the channel?

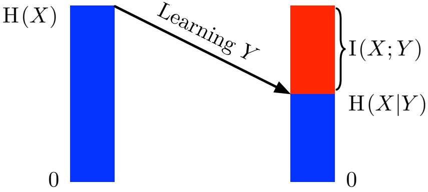
Answer: $\mathrm{H}(X) - \mathrm{H}(X|Y)$.

1 Mutual Information
- Definitions
- Properties

2 Noisy Channel Coding Theorem
- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Mutual information





Definition 1 (Mutual Information)

For a pair of jointly distributed (X, Y), the mutual information between X and Y is defined as

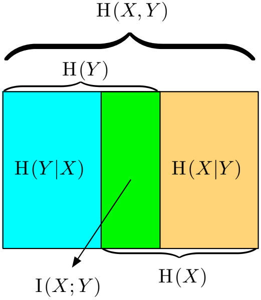
$$\mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X|Y).$$

Interpretation: Infer information about the channel input X from output Y.

Proposition 1 (An Identity)

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$
$$= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y).$$

pf: By chain rule: $\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - \mathrm{H}(Y)$. □



Note: Mutual information is symmetric, that is,

$$\mathrm{I}(X;Y) = \mathrm{I}(Y;X).$$

The mutual information between X and itself is equal to its entropy:

$$\mathrm{I}(X;X) = \mathrm{H}(X) \text{ since } \mathrm{H}(X|X) = 0.$$

Hence, the entropy is also called “self information” in some literatures.

Mutual information and information divergence

Proposition 2

For $(X, Y) \sim \mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X} = \mathsf{P}_Y \mathsf{P}_{X|Y}$*,*

$$\mathrm{I}(X;Y) = \mathsf{E}_{(X,Y)\sim \mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{P}_X(X)\mathsf{P}_Y(Y)}\right] = \mathrm{D}\left(\mathsf{P}_{X,Y} \| \mathsf{P}_X \times \mathsf{P}_Y\right),$$

where $\mathsf{P}_X \times \mathsf{P}_Y$ *is a product distribution of the two marginals* P_X *and* P_Y*.*

pf: A simple corollary of the identity in Proposition 1. □

Corollary 1 (Extremal Values of Mutual Information)

1. $\mathrm{I}(X;Y) \geq 0$*, with equality iff* X, Y *are independent.*
2. $\mathrm{I}(X;Y) \leq \mathrm{H}(X)$*, with equality iff* X *is a deterministic function of* Y*.*

pf: The proof of the first one is due to the fact that conditioning reduces entropy. The proof of the second one is due to $\mathrm{H}(X|Y) \geq 0$.

Interpretation: the mutual information between X and Y, $\mathrm{I}(X;Y)$ can also be viewed as a measure of the dependency between X and Y.

It is the divergence of the actual distribution $\mathsf{P}_{X,Y}$ from the independent distribution $\mathsf{P}_X \times \mathsf{P}_Y$.

- If X is determined by Y (highly dependent), $\mathrm{I}(X;Y)$ is maximized.
- If X is independent of Y (no dependency), $\mathrm{I}(X;Y) = 0$.

Exercise 1

Prove the following identity:

$$\mathrm{I}(X;Y) = \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{P}_Y \middle| \mathsf{P}_X\right) = \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_Y \middle| \mathsf{P}_X\right) - \mathrm{D}\left(\mathsf{P}_Y \| \mathsf{Q}_Y\right),$$

$\forall\, \mathsf{Q}_Y$ such that $\mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) < \infty$. Furthermore,

$$\mathrm{I}(X;Y) = \min_{\mathsf{Q}_Y : \mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) < \infty} \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_Y \middle| \mathsf{P}_X\right).$$

Example 1

Compute $\mathrm{I}(X_1;X_2)$ for $X_1, X_2 \in \{0,1\}$ with joint PMF

(x_1,x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}(x_1,x_2)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

sol: From the previous examples, we have

$$\begin{aligned}\mathrm{H}(X_1,X_2) = \log 3 + \tfrac{1}{3},\ \mathrm{H}(X_1) = \mathrm{H}(X_2) = 1,\\ \mathrm{H}(X_1|X_2) = \mathrm{H}(X_2|X_1) = \log 3 - \tfrac{2}{3}.\end{aligned}$$

Hence, $\mathrm{I}(X_1;X_2) = \mathrm{H}(X_1) - \mathrm{H}(X_1|X_2) = \frac{5}{3} - \log 3$.

Conditional mutual information

Definition 2 (Conditional Mutual Information)

For a tuple of jointly distributed r.v.'s (X, Y, Z), the mutual information between X and Y given Z is defined as

$$\mathrm{I}(X;Y|Z) := \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z).$$

Similar to the previous identity (Proposition 1), we have

$$\begin{aligned}\mathrm{I}(X;Y|Z) &= \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z) = \mathrm{H}(Y|Z) - \mathrm{H}(Y|X,Z) \\ &= \mathrm{H}(X|Z) + \mathrm{H}(Y|Z) - \mathrm{H}(X,Y|Z).\end{aligned}$$

Similar to Corollary 1, we have

1. $\mathrm{I}(X;Y|Z) \geq 0$, with equality iff X, Y are independent given Z, that is, $X - Z - Y$ forms a Markov chain.
2. $\mathrm{I}(X;Y|Z) \leq \mathrm{H}(X|Z)$, with equality iff X is a (deterministic) function of Y, Z.

1 Mutual Information

- Definitions
- Properties

2 Noisy Channel Coding Theorem

- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Chain rule for mutual information

Theorem 1 (Chain Rule for Mutual Information)

$$\mathrm{I}(X;Y^n) = \textstyle\sum_{i=1}^{n} \mathrm{I}\left(X;Y_i \middle| Y^{i-1}\right).$$

pf: Proved by definition and the chain rule for entropy. □

Exercise 2

Show that

$$\mathrm{I}(X;Z) \le \mathrm{I}(X;Y,Z) \quad \text{and} \quad \mathrm{I}(X;Y|Z) \le \mathrm{I}(X;Y,Z)\,.$$

Data processing inequality

Theorem 2 (Data Processing Inequality)

For a Markov chain $X - Y - Z$, that is, $\mathsf{P}_{X,Y,Z} = \mathsf{P}_X \mathsf{P}_{Y|X} \mathsf{P}_{Z|Y}$, we have

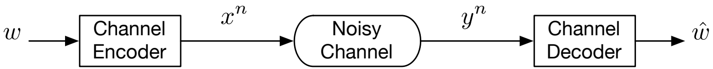
$$\mathrm{I}(X;Y) \ge \mathrm{I}(X;Z)\,.$$

Interpretation: $X - Y - Z$ says that the information of X that Z can provide is contained in Y. Hence, the amount of information of X that can be inferred by Z is not greater than that can be inferred by Y.

pf: Since $X - Y - Z$, we have $\mathrm{I}(X;Z|Y) = 0$. Hence,

$$\begin{aligned} &\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Y) + \mathrm{I}(X;Z|Y) = \mathrm{I}(X;Y) && (\because \mathrm{I}(X;Z|Y) = 0) \\ &\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y|Z) && \text{(Chain Rule)} \\ \Longrightarrow\ &\mathrm{I}(X;Y) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y|Z) \ge \mathrm{I}(X;Z)\,. \end{aligned}$$

□



Markov chains are common in communication systems. For example, in channel coding (without feedback), the message W, the channel input X^n, the channel output Y^n, and the decoded message $\hat{W}$ form a Markov chain $W - X^n - Y^n - \hat{W}$ (see the figure above).

Data processing inequality turns out to be crucial in obtaining impossibility results in information theory.

Exercise 3 (Functions of R.V.)

For $Z = g(Y)$ being a deterministic function of Y, show that

$$\mathrm{H}(Y) \ge \mathrm{H}(Z) \quad \text{and} \quad \mathrm{I}(X;Y) \ge \mathrm{I}(X;Z)\,.$$

Exercise 4

Show that

$$X_1 - X_2 - X_3 - X_4 \implies \mathrm{I}(X_1;X_4) \le \mathrm{I}(X_2;X_3)\,.$$

Example 2

Consider two random variables $X_1, X_2 \in \{0, 1\}$ with the same joint PMF as that in Example 1. Let $X_3 = X_2 \oplus Z$, where $Z \sim \mathrm{Ber}(p)$ and Z is independent of (X_1, X_2).

1. Compute $\mathrm{I}(X_1; X_3)$ and $\mathrm{I}(X_1; X_2|X_3)$.
2. Show that $X_1 - X_2 - X_3$ forms a Markov chain.
3. Verify the data processing inequality $\mathrm{I}(X_1; X_2) \geq \mathrm{I}(X_1; X_3)$.

sol:

(x_1, x_2, x_3)	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$
$\mathsf{P}(x_1, x_2, x_3)$	$\frac{1}{6}(1-p)$	$\frac{1}{6}p$	$\frac{1}{3}p$	$\frac{1}{3}(1-p)$
(x_1, x_2, x_3)	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$\mathsf{P}(x_1, x_2, x_3)$	$\frac{1}{3}(1-p)$	$\frac{1}{3}p$	$\frac{1}{6}p$	$\frac{1}{6}(1-p)$

Then it is straightforward to compute mutual informations and verify the Markov chain $X_1 - X_2 - X_3$.

Conditioning reduces mutual information?

Sometimes yes, sometimes no.

Proposition 3 (Conditioning May Decrease Mutual Information)

For a Markov chain $X - Y - Z$, we have $\mathrm{I}(X;Y) \ge \mathrm{I}(X;Y|Z)$.

pf: The same argument as that of the data processing inequality. □

Example 3 (Conditioning May Increase Mutual Information)

Let X and Y be i.i.d. $\mathrm{Ber}(\frac{1}{2})$ random variables, and $Z = X \oplus Y$. Evaluate $\mathrm{I}(X;Y|Z)$ and show that $\mathrm{I}(X;Y|Z) > \mathrm{I}(X;Y)$.

sol:
$$\begin{aligned}\mathrm{I}(X;Y|Z) &= \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z) = \mathrm{H}(X|Z) - \mathrm{H}(X|Y,X \oplus Y)\\ &= \mathrm{H}(X|Z) - \mathrm{H}(X|Y,X) = \mathrm{H}(X|Z) = \mathrm{H}(X) = 1.\end{aligned}$$

(note that X and Z are independent)

On the other hand, $\mathrm{I}(X;Y) = 0$. Hence, $1 = \mathrm{I}(X;Y|Z) > \mathrm{I}(X;Y) = 0$.

Convexity and concavity of mutual information

Theorem 3

Let $(X, Y) \sim \mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X}$*.*

1. *With* $\mathsf{P}_{Y|X}$ *fixed,* $\mathrm{I}(X;Y)$ *is a concave function of* P_X*.*
2. *With* P_X *fixed,* $\mathrm{I}(X;Y)$ *is a convex function of* $\mathsf{P}_{Y|X}$*.*

pf: For the first property, note that

$$\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) = \mathrm{H}(\mathsf{P}_Y) - \underbrace{\textstyle\sum_{x \in \mathcal{X}} \mathsf{P}_X(x) \mathrm{H}(Y|X=x)}_{\text{linear in } \mathsf{P}_X}.$$

It suffices to show that $\mathrm{H}(\mathsf{P}_Y)$ is concave in P_X.

P_Y is a linear function of P_X and $\mathrm{H}(\mathsf{P}_Y)$ is a concave function of P_Y. As a result, $\mathrm{H}(\mathsf{P}_Y)$ is concave in P_X, which completes the proof of concavity.

For the convexity, let us consider two conditional distributions $\mathsf{P}^{(0)}_{Y|X}, \mathsf{P}^{(1)}_{Y|X}$, and random variables X, Y, Θ such that $\mathsf{P}_{\Theta,X,Y} = \mathsf{P}_X \mathsf{P}_\Theta \mathsf{P}_{Y|X,\Theta}$, $\Theta \sim \mathrm{Ber}(\lambda)$, and $\mathsf{P}_{Y|X,\Theta}(y\,|\,x,\theta) = \mathsf{P}^{(\theta)}_{Y|X}(y\,|\,x)$, $\forall\, y \in \mathcal{Y},\ x \in \mathcal{X},\ \theta \in \{0,1\}$.

Hence, $\Theta \perp\!\!\!\perp X$ and $\mathsf{P}_{Y|X} = (1-\lambda)\mathsf{P}^{(0)}_{Y|X} + \lambda \mathsf{P}^{(1)}_{Y|X}$.

By the basic properties of mutual information and entropy,

$$\mathrm{I}(X;Y) \le \mathrm{I}(X;Y,\Theta) = \mathrm{I}(X;Y|\Theta) + \mathrm{I}(X;\Theta)\,.$$

Note that $\Theta \perp\!\!\!\perp X \implies \mathrm{I}(X;\Theta) = 0$. Also note that

- $\mathrm{I}(X;Y) = \mathrm{I}(X;Y)\,\big|_{(X,Y)\sim \mathsf{P}_X\left\{(1-\lambda)\mathsf{P}^{(0)}_{Y|X} + \lambda \mathsf{P}^{(1)}_{Y|X}\right\}}$.
- $\mathrm{I}(X;Y|\Theta) = (1-\lambda)\mathrm{I}(X;Y|\Theta=0) + \lambda \mathrm{I}(X;Y|\Theta=1)$ and

$$\mathrm{I}(X;Y|\Theta=\theta) = \mathrm{I}(X;Y)\,\big|_{(X,Y)\sim \mathsf{P}_X \mathsf{P}^{(\theta)}_{Y|X}},\ \theta \in \{0,1\}.$$

The proof is complete. □

Summary: mutual Information

- $\mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X|Y)$ measures the amount of information of X contained in Y (and vice versa). $\mathrm{I}(X;Y) = \mathrm{D}(\mathsf{P}_{X,Y} \| \mathsf{P}_X \times \mathsf{P}_Y)$ also tells the level of dependency between X and Y.
- $\mathrm{I}(X;Y|Z) := \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z)$ measures the amount of information of X in Y given Z.
- Nonnegative; Concave function of P_X with fixed $\mathsf{P}_{Y|X}$; Convex function of $\mathsf{P}_{Y|X}$ with fixed P_X.
- Chain rule: $\mathrm{I}(X;Y^n) = \sum_{i=1}^n \mathrm{I}\left(X;Y_i \middle| Y^{i-1}\right)$.
- Conditioning may reduce MI: $X - Y - Z \implies \mathrm{I}(X;Y|Z) \le \mathrm{I}(X;Y)$.
- Data processing decreases MI: $X - Y - Z \implies \mathrm{I}(X;Y) \ge \mathrm{I}(X;Z)$.

Next: noisy channel coding theorem.

1 Mutual Information

- Definitions
- Properties

2 Noisy Channel Coding Theorem

- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Discrete memoryless channel (DMC)

We focus on a main class of channels, *discrete memoryless channels* (DMC), where the perturbation experienced in the channel is memoryless.

Definition 3 (Discrete Memoryless Channel)

A discrete channel $\left(\mathcal{X}, \left\{\mathsf{P}_{Y_i|X^i,Y^{i-1}} \,\middle|\, i \in \mathbb{N}\right\}, \mathcal{Y}\right)$ is *memoryless* if $\forall\, i \in \mathbb{N}$,

$$\mathsf{P}_{Y_i|X^i,Y^{i-1}}(y_i|x^i, y^{i-1}) = \mathsf{P}_{Y|X}(y_i|x_i).$$

In other words, $Y_i - X_i - \left(X^{i-1}, Y^{i-1}\right)$. Here the conditional PMF $\mathsf{P}_{Y|X}$ is called the *channel law* or *channel transition function*.

Note: The condition in the above definition is **NOT** sufficient to specify

$$\mathsf{P}_{Y^n|X^n}(y^n|x^n),$$

the stochastic relationship between the channel input (codeword) x^n and the channel output y^n.

Why? Because

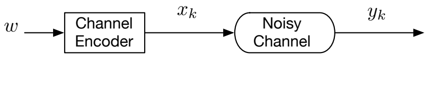
$$\begin{aligned}
\mathsf{P}(y^n|x^n) &= \frac{\mathsf{P}(x^n, y^n)}{\mathsf{P}(x^n)} \\
\mathsf{P}(x^n, y^n) &= \prod_{i=1}^{n} \mathsf{P}(x_i, y_i|x^{i-1}, y^{i-1}) \\
&= \prod_{i=1}^{n} \mathsf{P}(y_i|x^i, y^{i-1})\mathsf{P}(x_i|x^{i-1}, y^{i-1})
\end{aligned}$$

Hence, we need to further specify $\left\{\mathsf{P}(x_i|x^{i-1}, y^{i-1}) \,\middle|\, i \in \mathbb{N}\right\}$, which cannot be obtained from $\mathsf{P}(x^n)$ alone.

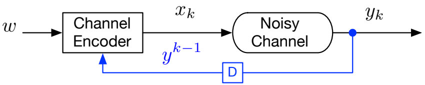
Interpretation: $\left\{\mathsf{P}(x_i|x^{i-1}, y^{i-1}) \,\middle|\, i \in \mathbb{N}\right\}$ is induced by the encoding function, which implies that the encoder can potentially make use of the past channel output, i.e., *feedback*.

Let us first focus on the channel coding problem without feedback.

DMC without feedback



No Feedback



With Feedback

Suppose in the system, the encoder has no knowledge about the realization of the channel output, then, $\mathsf{P}(x_i|x^{i-1}, y^{i-1}) = \mathsf{P}(x_i|x^{i-1})$ for all $i \in \mathbb{N}$.
$\implies$ It is said that the channel has no feedback.

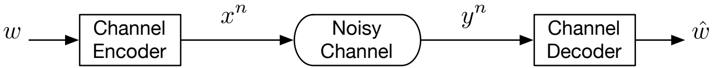
In this case, specifying $\left\{\mathsf{P}(y_i|x^i, y^{i-1}) \,\middle|\, i \in \mathbb{N}\right\}$ suffices to specify $\mathsf{P}(y^n|x^n)$.

Proposition 4 (DMC without Feedback)

For a DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$ *without feedback,*

$$\mathsf{P}(y^n|x^n) = \prod_{i=1}^n \mathsf{P}_{Y|X}(y_i|x_i).$$

Channel coding without feedback: problem setup



1. An (n,k) channel code consists of
   - an encoding function (encoder) $\mathrm{enc}_n : \mathcal{W} \equiv \{1,2,...,2^k\} \to \mathcal{X}^n$ that maps each message w to a length n codeword x^n.
   - a decoding function (decoder) $\mathrm{dec}_n : \mathcal{Y}^n \to \mathcal{W}$ that maps a channel output sequence y^n to a reconstructed message $\hat{w}$
2. An (n,k) code with $\mathsf{P}_{\mathsf{e}}^{(n)} := \Pr\left\{W \neq \hat{W}\right\} \le \epsilon$ is called an (n,k,ϵ) code.
3. Let $k^*(n,\epsilon)$ denote the largest k such that there exists an (n,k,ϵ) code. The ϵ-capacity $\mathrm{C}(\epsilon) := \lim\limits_{n\to\infty} \frac{1}{n} k^*(n,\epsilon)$ if the limit exists.

   Meanwhile, the Shannon capacity $\mathrm{C} := \lim_{\epsilon\downarrow 0} \mathrm{C}(\epsilon)$.

Channel coding theorem for DMC

Theorem 4 (Channel Coding Theorem for DMC without Feedback)

The ϵ-capacity $\mathrm{C}(\epsilon)$ of the DMC $\left(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y}\right)$ without feedback is characterized as follows:

$$\mathrm{C}(\epsilon) = \mathrm{C}^{\mathrm{I}} := \max_{\mathsf{P}_X \in \mathcal{P}(\mathcal{X})} \mathrm{I}(X;Y) \quad \forall\, \epsilon \in (0,1). \tag{1}$$

Here $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions over the alphabet $\mathcal{X}$, that is, $\mathcal{P}(\mathcal{X}) \equiv \mathcal{P}_{|\mathcal{X}|}$ is a probability simplex in a $|\mathcal{X}|$-dimensional space.

Remarks:

- Capacity formula (1) is intuitive: $\mathrm{I}(X;Y)$ is just the amount of information about the channel input X that one can infer from the output Y.
- The maximization over $\mathsf{P}_X \in \mathcal{P}(\mathcal{X})$ stands for choosing the best possible input distribution so that the amount of information transfer is maximized.
- As a corollary, the Shannon capacity $\mathrm{C} = \max_{\mathsf{P}_X} \mathrm{I}(X;Y)$ as well.

Shannon capacity and ϵ-capacity: some remarks

In the previous formulation, definition of the Shannon capacity relies on the existence of the ϵ-capacity, that is, the limit $\lim_{n\to\infty} \frac{1}{n} k^*(n,\epsilon)$.

In fact, capacity can be defined alternatively *without relying on the existence of the limit*. For example, in Chapter 19 of Polyanskiy and Wu, ϵ-capacity is defined as the limit inferior $\liminf_{n\to\infty} \frac{1}{n} k^*(n,\epsilon)$ and hence always exist.

Meanwhile, historically, Shannon capacity was investigated first, which can be defined equivalently as follows:

> *A rate R is said to be ϵ-achievable (achievable) if there exist a sequence of $(n, \lceil n\mathrm{R} \rceil)$ codes such that $\mathsf{P}_{\mathsf{e}}^{(n)} \le \epsilon$ when n is large enough ($\mathsf{P}_{\mathsf{e}}^{(n)} \to 0$ as $n \to \infty$).*
>
> *The ϵ-capacity is defined as $\mathrm{C}(\epsilon) := \sup\{\mathrm{R} \,|\, \mathrm{R} : \epsilon\text{-achievable}\}$.*
>
> *The Shannon capacity is defined as $\mathrm{C} := \sup\{\mathrm{R} \,|\, \mathrm{R} : \text{achievable}\}$.*

A weaker version of channel coding theorem (Theorem 4) with $\mathrm{C}(\epsilon)$ replaced by C was proved (converse: Fano 1952; achievability: Feinstein 1954) first. Later, the converse of the stronger version was proved by Wolfowitz in 1957.

Note: the converse of the stronger version (termed *strong converse* in the literature) equivalently says that if $\mathrm{R} > \mathrm{C}^{\mathrm{I}}$, $\mathsf{P}_{\mathsf{e}}^{(n)} \to 1$, while the weaker version (*weak converse*) only guarantees $\mathsf{P}_{\mathsf{e}}^{(n)} \not\to 0$.

In this lecture, we will prove the weak converse using Fano's inequality, and demonstrate how to extend the strong converse argument of BSC capacity only when time permits.

Roadmap

1. **Capacity computation**:

   Give some simple examples of DMCs to show how to compute capacity.

2. **Weak converse**:

   Prove that $\forall\, \mathrm{R} > \mathrm{C}^{\mathrm{I}}$, it is impossible to have vanishing error probability, that is, $\mathsf{P}_{\mathsf{e}}^{(n)} \not\to 0$ as $n \to \infty$.

3. **Achievability**:

   Prove that $\forall\, \mathrm{R} < \mathrm{C}^{\mathrm{I}}$, there exist a sequence of encoding/decoding schemes such that the error probability vanishes, that is, $\mathsf{P}_{\mathsf{e}}^{(n)} \to 0$ as $n \to \infty$, based on a probabilistic argument called random coding.

1 Mutual Information
- Definitions
- Properties

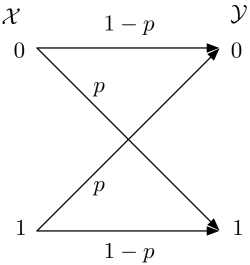
2 Noisy Channel Coding Theorem
- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Binary symmetric channel

A binary symmetric channel (BSC) consists of

- Binary input/output $\mathcal{X} = \mathcal{Y} = \{0, 1\}$.
- Channel law $\mathsf{P}_{Y|X} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$.

The capacity of BSC is $\mathrm{C}_{\mathrm{BSC}} = 1 - \mathsf{H}_\mathsf{b}(p)$.



To compute BSC capacity, observe $\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$, and

- $\mathrm{H}(Y|X=0) = \mathrm{H}(Y|X=1) = \mathsf{H}_\mathsf{b}(p) \implies \mathrm{H}(Y|X) = \mathsf{H}_\mathsf{b}(p)$.
- $\mathrm{H}(Y) \le \log 2 = 1$, with equality iff Y is uniform.

Question: Is it possible to choose a P_X such that Y is uniform?

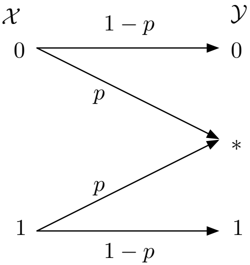
Ans: Yes – choose X to be uniform $\implies \mathrm{C} = \max_{\mathsf{P}_X} \mathrm{I}(X;Y) = 1 - \mathsf{H}_\mathsf{b}(p)$.

Binary erasure channel

A binary erasure channel (BEC) consists of

- Binary input $\mathcal{X} = \{0,1\}$ and output with erasure $\mathcal{Y} = \{0,1,*\}$.
- Channel law $\mathsf{P}_{Y|X} = \begin{bmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{bmatrix}$.

The capacity of BEC is $\mathrm{C}_{\mathrm{BEC}} = 1 - p$.

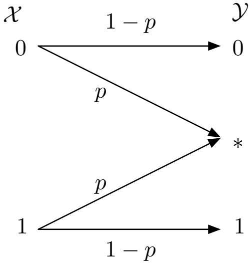


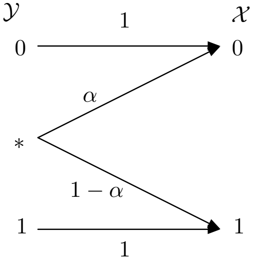
Suppose we begin with $\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$. Then,

- $\mathrm{H}(Y|X=0) = \mathrm{H}(Y|X=1) = \mathsf{H}_{\mathsf{b}}(p) \implies \mathrm{H}(Y|X) = \mathsf{H}_{\mathsf{b}}(p)$.
- $\mathrm{H}(Y) \le \log 3$, with equality iff Y is uniform.

Question: Is it possible to choose a P_X such that Y is uniform?

Ans: No. So, we cannot say that $\max_{\mathsf{P}_X} \mathrm{H}(Y) = \log 3$.





Instead, we can start with $\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y)$. Then, we have the *reverse* channel law $\mathsf{P}_{X|Y} = \begin{bmatrix} 1 & 0 \\ \alpha & 1-\alpha \\ 0 & 1 \end{bmatrix}$, where $\alpha := \Pr\{X = 0\} \equiv \mathsf{P}_X(0)$.

- $\mathrm{H}(X|Y=0) = \mathrm{H}(X|Y=1) = 0,\ \mathrm{H}(X|Y=*) = \mathsf{H_b}(\alpha) = \mathrm{H}(X)$ $\implies \mathrm{H}(X|Y) = \Pr\{Y = *\}\,\mathrm{H}(X) = p\mathrm{H}(X)$.
- $\mathrm{H}(X) \le 1$, with equality iff X is uniform.

Hence, $\mathrm{C}_{\mathrm{BEC}} = \max_{\mathsf{P}_X} (1-p)\,\mathrm{H}(X) = 1-p$.

Erasure channel

We can generalize BEC to the following erasure channel:

- Input $\mathcal{X}$, output $\mathcal{Y} = \mathcal{X} \cup \{*\}$.
- Channel law $\mathsf{P}_{Y|X} = \begin{cases} 1-p, & y = x \\ p, & y = * \\ 0, & \text{otherwise} \end{cases}$

A motivation for this model is from networking, where the erasure $*$ models the “packet drop”.

Exercise 5

Show that the capacity of the erasure channel is

$$\mathrm{C}_{\mathrm{EC}} = (1-p)\log|\mathcal{X}|.$$

Symmetric channel

In computing the capacity of BSC, we observe that

1. $\mathrm{H}(Y|X) = \mathsf{H_b}(p)$ regardless of P_X. Why? Because all rows of $\mathsf{P}_{Y|X}$ are permutations of a same probability vector $(p, 1-p)$.
2. $\mathrm{H}(Y) = \log|\mathcal{Y}|$ can be attained, that is, Y can be made uniform by choosing X to be uniform. Why? Because all columns of $\mathsf{P}_{Y|X}$, have the same sum $\sum_x \mathsf{P}_{Y|X}(y|x)$.

Definition 4 (Symmetric Channel)

A symmetric channel is a channel with channel law $\mathsf{P}_{Y|X}$ satisfying (1) all rows of $\mathsf{P}_{Y|X}$ are permutations of a same probability vector $\boldsymbol{p}$, and (2) all columns of $\mathsf{P}_{Y|X}$, have the same sum $\sum_x \mathsf{P}_{Y|X}(y|x)$.

Exercise 6

Show that the capacity of a symmetric channel is $\log|\mathcal{Y}| - \mathrm{H}(\boldsymbol{p})$.

Computing DMC capacity by convex optimization

For a DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$, since $\mathrm{I}(X;Y)$ is a concave function of P_X for fixed $\mathsf{P}_{Y|X}$, to find the channel capacity of this DMC, we only need to solve a convex program in the capacity formula (1).

There are hence several avenues towards computing the channel capacity.

1. Try to solve it analytically, either by inspection (using structures of the channel law $\mathsf{P}_{Y|X}$), or write down the KKT conditions to find the optimal solution.
2. Call for fast algorithms to solve convex programs over a probability simplex (*conditional gradient descent* for example).
3. Leverage the problem structure and devise specific iterative algorithms.

In the following, let us introduce a widely used algorithm (Blahut-Arimoto algorithm) due to Blahut and Arimoto, which belongs to the third category.

Blahut[1]-Arimoto[2] algorithm

Proposition 5 (A variational characterization of mutual information)

Let $(X, Y) \sim \mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X} = \mathsf{P}_Y \mathsf{P}_{X|Y}$*. Then,*

$$\mathrm{I}(X;Y) = \max_{\mathsf{Q}_{X|Y}} \left\{ \mathsf{E}_{\mathsf{P}_{X,Y}} \left[\log \frac{\mathsf{Q}_{X|Y}(X|Y)}{\mathsf{P}_X(X)} \right] \right\},$$

with $\mathsf{P}_{X|Y}$ *being the maximizer.*

As a result, for a DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$, its capacity can be written as a double maximization as follows:

$$\mathrm{C} = \max_{\mathsf{P}} \max_{\mathsf{Q}} \left\{ \sum_x \sum_y \mathsf{P}(x) \mathsf{P}_{Y|X}(y|x) \log \frac{\mathsf{Q}(x|y)}{\mathsf{P}(x)} \right\}. \tag{2}$$

[1] R. E. Blahut, “Computation of channel capacity and rate distortion functions,” IEEE Transactions on Information Theory, vol. 18, no. 4, 1972.

[2] S. Arimoto, “An algorithm for calculating the capacity of arbitrary discrete memoryless channels,” IEEE Transactions on Information Theory, vol. 18, no. 1, 1972.

pf: Goal: show $\mathrm{I}(X;Y) \geq \mathsf{E}_{\mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{Q}_{X|Y}(X|Y)}{\mathsf{P}_X(X)}\right]$ with equality if $\mathsf{Q}_{X|Y} = \mathsf{P}_{X|Y}$.

Note that $\mathrm{I}(X;Y) = \mathsf{E}_{\mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{P}_{X|Y}(X|Y)}{\mathsf{P}_X(X)}\right]$, and hence

$$\text{RHS} - \text{LHS} = \mathsf{E}_{\mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{Q}_{X|Y}(X|Y)}{\mathsf{P}_{X|Y}(X|Y)}\right] \leq \log\left(\mathsf{E}_{\mathsf{P}_{X,Y}}\left[\frac{\mathsf{Q}_{X|Y}(X|Y)}{\mathsf{P}_{X|Y}(X|Y)}\right]\right),$$

where the last inequality is due to concavity of $\log(\cdot)$ and Jensen's inequality. Hence, it holds with equality if $\mathsf{Q}_{X|Y}(x|y) = \mathsf{P}_{X|Y}(x|y)$ for all x, y.

Finally, note that

$$\begin{aligned}\mathsf{E}_{\mathsf{P}_{X,Y}}\left[\frac{\mathsf{Q}_{X|Y}(X|Y)}{\mathsf{P}_{X|Y}(X|Y)}\right] &= \sum_y \mathsf{P}_Y(y) \sum_x \mathsf{P}_{X|Y}(x|y) \frac{\mathsf{Q}_{X|Y}(x|y)}{\mathsf{P}_{X|Y}(x|y)} \\ &= \sum_y \mathsf{P}_Y(y) \sum_x \mathsf{Q}_{X|Y}(x|y) \leq 1,\end{aligned}$$

with equality if $\mathsf{supp}_{\mathsf{Q}_{X|Y}(\cdot|y)} = \mathsf{supp}_{\mathsf{P}_{X|Y}(\cdot|y)}$ for all $y \in \mathsf{supp}_{\mathsf{P}_Y}$.

The proof is complete. □

(2) says that the capacity of a DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$ can be expressed as a double maximum of a *bounded* concave function $h(\mathsf{P}, \mathsf{Q})$, where

$$h(\mathsf{P}, \mathsf{Q}) := \sum_x \sum_y \mathsf{P}(x)\mathsf{P}_{Y|X}(y|x) \log \frac{\mathsf{Q}(x|y)}{\mathsf{P}(x)},$$

jointly concave in P and Q. Below is a natural alternating algorithm which gives a strictly increasing sequence of h's[3].

0 Initialize with some $\mathsf{P}^{(0)} \in \mathcal{P}_+(\mathcal{X})$. Set $l = 1$.

1 $\mathsf{Q}^{(l)} \leftarrow \arg\max_{\mathsf{Q}} h(\mathsf{P}^{(l-1)}, \mathsf{Q})$. By Proposition 5,

$$\mathsf{Q}^{(l)}(x|y) = \left(\mathsf{P}^{(l-1)}(x)\mathsf{P}_{Y|X}(y|x)\right) \Big/ \left(\sum_{x\in\mathcal{X}} \mathsf{P}^{(l-1)}(x)\mathsf{P}_{Y|X}(y|x)\right).$$

2 $\mathsf{P}^{(l)} \leftarrow \arg\max_{\mathsf{P}\in\mathcal{P}_+(\mathcal{X})} h(\mathsf{P}, \mathsf{Q}^{(l)})$. By solving the KKT conditions, we get

$$\mathsf{P}^{(l)}(x) = \left(\prod_y \mathsf{Q}^{(l)}(x|y)^{\mathsf{P}_{Y|X}(y|x)}\right) \Big/ \left(\sum_{x\in\mathcal{X}} \prod_y \mathsf{Q}^{(l)}(x|y)^{\mathsf{P}_{Y|X}(y|x)}\right).$$

3 $l \leftarrow l + 1$ and return to Step 1.

[3] Check Ch. 9 of R. Yeung, Information Theory and Network Coding, Springer, 2008, for convergence guarantee.

1 Mutual Information
- Definitions
- Properties

2 Noisy Channel Coding Theorem
- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Fano's inequality

Lemma 1 (Fano's Inequality)

$\mathrm{H}(U|V) \le \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e} \log|\mathcal{U}|$, *where* $\mathsf{P_e} := \mathsf{Pr}\{U \neq V\}$.

pf: Let $E := \mathbb{1}\{U \neq V\}$, the indicator function of $\{U \neq V\}$. $E \sim \mathrm{Ber}(\mathsf{P_e})$.

Using chain rule and the non-negativity of conditional entropy, we have

$$\mathrm{H}(U|V) \le \mathrm{H}(U,E|V) = \mathrm{H}(E|V) + \mathrm{H}(U|V,E)\,.$$

Note that $\mathrm{H}(E|V) \le \mathrm{H}(E) = \mathsf{H_b}(\mathsf{P_e})$, and

$$\mathrm{H}(U|V,E) = \underbrace{\mathsf{Pr}\{E=1\}}_{=\mathsf{P_e}} \underbrace{\mathrm{H}(U|V,E=1)}_{\le \log|\mathcal{U}|} + \mathsf{Pr}\{E=0\} \underbrace{\mathrm{H}(U|V,E=0)}_{=0,\ \because U=V}$$

Hence, $\mathrm{H}(U|V) \le \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e} \log|\mathcal{U}|$. □

Corollary 2 (Lower Bound on Error Probability)

$$\mathsf{P}_\mathsf{e} \ge \frac{\mathrm{H}(U|V) - 1}{\log|\mathcal{U}|}.$$

pf: From Lemma 1 and $\mathsf{H_b}(\mathsf{P_e}) \le 1$, we have

$$\mathrm{H}(U|V) \le \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e}\log|\mathcal{U}| \le 1 + \mathsf{P_e}\log|\mathcal{U}|.$$ □

Exercise 7

Show that Lemma 1 can be sharpened as follows

$$\mathrm{H}(U|V) \le \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e}\log(|\mathcal{U}| - 1),$$

if U, V both take values in $\mathcal{U}$.

Fano's inequality is widely used in proving converse parts (impossibility results) of coding theorems in information theory.

Proof of the (weak) converse

We would like to show that for every sequence of $(n, \lceil n\mathrm{R} \rceil)$ codes such that $\mathsf{P}_{\mathsf{e}}^{(n)} \to 0$ as $n \to \infty$, the rate $\mathrm{R} \le \mathrm{C}^{\mathrm{I}} := \max_{\mathsf{P}_X} \mathrm{I}(X;Y)$.

pf: Let $k = \lceil n\mathrm{R} \rceil$. Note that $W \sim \mathrm{Unif}\{1, 2, ..., 2^k\}$ and hence $k = \mathrm{H}(W)$.

$$n\mathrm{R} \le \lceil n\mathrm{R} \rceil = \mathrm{H}(W) = \mathrm{I}\left(W; \hat{W}\right) + \mathrm{H}\left(W \middle| \hat{W}\right) \tag{3}$$

$$\le \mathrm{I}(W; Y^n) + \left(1 + \mathsf{P}_{\mathsf{e}}^{(n)} \log\left(2^k\right)\right) \tag{4}$$

- (3) is due to $\lceil n\mathrm{R} \rceil \ge n\mathrm{R}$ and the chain rule.
- (4) is due to $W - Y^n - \hat{W}$ and Fano's inequality ($|\mathcal{W}| = 2^k$).

Expanding $\mathrm{I}(W; Y^n)$ in (4) by the chain rule and using the fact that $\lceil n\mathrm{R} \rceil \le n\mathrm{R} + 1$, we reach the following upper bound on R:

$$\mathrm{R} \le \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left(W; Y_i \middle| Y^{i-1}\right) + \frac{1}{n} \left(1 + \mathsf{P}_{\mathsf{e}}^{(n)} (n\mathrm{R} + 1)\right)$$

Set $\varepsilon_n := \frac{1}{n}\left(1 + \mathsf{P}_\mathsf{e}^{(n)}\left(n\mathrm{R}+1\right)\right)$. $\varepsilon_n \to 0$ as $n \to \infty$ because $\lim_{n\to\infty} \mathsf{P}_\mathsf{e}^{(n)} = 0$.

Next: relate $\sum_{i=1}^{n} \mathrm{I}\left(W; Y_i \middle| Y^{i-1}\right)$ to $\mathrm{I}\left(X; Y\right)$:

$$\mathrm{I}\left(W; Y_i \middle| Y^{i-1}\right) \le \mathrm{I}\left(W, Y^{i-1}; Y_i\right) \le \mathrm{I}\left(W, Y^{i-1}, X_i; Y_i\right) \overset{(*)}{=} \mathrm{I}\left(X_i; Y_i\right).$$

Here $(*)$ is due to the property of a DMC:

$$\begin{aligned} &\mathsf{P}(y_i|x^i, y^{i-1}, w) = \mathsf{P}(y_i|x^i, y^{i-1}) = \mathsf{P}(y_i|x_i) \\ \Longrightarrow\ & Y_i - X_i - \left(W, X^{i-1}, Y^{i-1}\right) \\ \Longrightarrow\ & Y_i - X_i - \left(W, Y^{i-1}\right). \end{aligned}$$

As a result, $\forall\, i$, $\mathrm{I}\left(W; Y_i \middle| Y^{i-1}\right) \le \mathrm{I}\left(X_i; Y_i\right) \le \max_{\mathsf{P}_X} \mathrm{I}\left(X; Y\right) =: \mathrm{C}^{\mathrm{I}}$.

Hence, we have $\forall\, n \in \mathbb{N}$,

$$\mathrm{R} \le \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\left(W; Y_i \middle| Y^{i-1}\right) + \varepsilon_n \le \mathrm{C}^{\mathrm{I}} + \varepsilon_n.$$

Taking $n \to \infty$, we conclude that $\mathrm{R} \le \mathrm{C}^{\mathrm{I}}$ if it is achievable. □

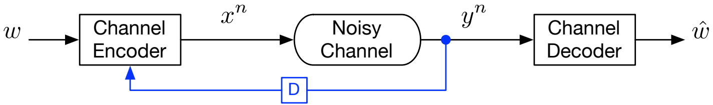
1 Mutual Information

- Definitions
- Properties

2 Noisy Channel Coding Theorem

- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Channel coding with feedback: problem setup



1. An (n, k) channel code consists of
   - an encoding function (encoder) $\text{enc}_n : \mathcal{W} \times \mathcal{Y}^{n-1} \to \mathcal{X}^n$ that maps each message w to a length n codeword x^n, and the i-th symbol x_i is a function of $\left(w, y^{i-1}\right)$ for all $i \in \{1, ..., n\}$.
   - a decoding function (decoder) $\text{dec}_n : \mathcal{Y}^n \to \mathcal{W}$ that maps a channel output sequence y^n to a reconstructed message $\hat{w}$.
2. The error probability is defined as $\mathsf{P}_{\mathsf{e}}^{(n)} := \Pr\left\{W \neq \hat{W}\right\}$.
3. A rate R is said to be *achievable* if there exist a sequence of $(n, \lceil n\mathrm{R} \rceil)$ codes such that $\mathsf{P}_{\mathsf{e}}^{(n)} \to 0$ as $n \to \infty$.

   The Shannon capacity is defined as $\mathrm{C} := \sup\{\mathrm{R} \mid \mathrm{R} : \text{achievable}\}$.

Feedback capacity

Theorem 5 (Channel Coding Theorem for DMC with Feedback)

The capacity of the DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$ *with feedback is given by* (1), *the same as that without feedback. In other words, feedback does not increase the channel capacity for DMC.*

Proof of the weaker version (feedback does not improve Shannon capacity) is immediate because in the converse proof of channel coding theorem without feedback, we do not make use of the assumption that there is no feedback. In other words, the proof is identical even with feedback.

Strong converse remains to hold (also due to Wolfowitz), that is, feedback does not improve ϵ-capacity either.

1 Mutual Information

- Definitions
- Properties

2 Noisy Channel Coding Theorem

- Channel Capacity
- Proof of Weak Converse
- Feedback Capacity
- Proof of Achievability
- Source-Channel Separation

Overview

To prove the achievability of Theorem 4, need to show the statement below:

$$\boxed{\forall\,\mathrm{R} < \mathrm{C^I}, \mathrm{R} \ge 0,\ \exists \text{ a sequence of } (n, \lceil n\mathrm{R}\rceil)\text{codes such that } \lim_{n\to\infty} \mathsf{P}_{\mathsf{e}}^{(n)} = 0.}$$

Reflection: there are two possible ways to prove the **existence** of objects satisfying some desirable properties:

1. Explicitly construct an object and show that the properties hold.

   (Feinstein's achievability proof)

2. Assume that no objects satisfy the properties, and show contradiction.

   (Shannon's achievability proof)

The achievability proof of the channel coding theorem presented in this lecture (and in most information theory textbooks) is of the second kind, and in fact belongs to the so-called **probabilistic method**.

Coding over a noisy channel

Let us set up a few notations related to coding over a noisy channel.

1. Codebook $\mathcal{C} = \left\{x^n(1), x^n(2), ..., x^n(2^k)\right\}$ consists of the 2^k *codewords* and is the range of the encoding function enc_n, that is, $\mathcal{C} = \mathrm{enc}_n(\mathcal{W})$.
2. Error probability of message w: $\lambda_w := \Pr\{\hat{W} \neq W \,|\, W = w\}$.

   The average error probability $\mathsf{P}_{\mathsf{e}}^{(n)} = \frac{1}{2^k}\sum_{w=1}^{2^k} \lambda_w$ depends on the decoding algorithm, the channel law, and the codebook $\mathcal{C}$.
3. Maximum likelihood decoder is the optimal decoder minimizing $\mathsf{P}_{\mathsf{e}}^{(n)}$ when the *random* message follows uniform prior:

$$\hat{w}_{\mathrm{ML}}(y^n) = \arg\max_{w \in \{1,...,2^k\}} \mathsf{P}(y^n | x^n(w)).$$

In principle, one can derive the ML decoding rule and compute $\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(\mathcal{C})$ for a given codebook $\mathcal{C}$.

However, there are some challenges toward proving the channel coding theorem with this approach.

Challenges and work-arounds

First, for a given codebook $\mathcal{C}$, the expression of $\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(\mathcal{C})$ is intractable in general, and it is hard to obtain insights regarding the asymptotic behaviors (because symbols of $X^n \equiv x^n(W) \sim \mathrm{Unif}(\mathcal{C})$ is correlated over time).

Second, it is unclear how to construct good codebooks with guarantee of vanishing error probability.

In summary, to prove the achievability part of the channel coding theorem, there are two main challenges we shall overcome:

1. **How to show the existence of good codebooks?**

   Circumvent the issue of explicit construction with a random coding argument (a kind of the probabilistic method)

2. **How to analyze the error probability?**

   Circumvent the issue of ML decoding error analysis with a suboptimal decoder and derive upper bounds on its error probability.

Proof program

1. **Random Codebook Generation**:

   Generate an ensemble of codebooks according to certain probability distribution. Hence, the codebook becomes a random object.

   Represent the codebook $\mathcal{C} = \left\{x^n(1), x^n(2), ..., x^n(2^k)\right\}$ as a $2^k \times n$ matrix $\mathbf{c}$, where each row corresponds to a codeword:

$$\mathbf{c} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^k) & x_2(2^k) & \cdots & x_n(2^k) \end{bmatrix}.$$

2. **Error Probability Analysis**:

   Goal: Show that $\lim_{n\to\infty} \mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(\mathbf{C})\right] = 0$ and conclude that there exists a codebook $\mathbf{c}$ such that error probability $\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(\mathbf{c}) \to 0$ as $n \to \infty$.

   Moreover, to simplify analysis, we shall introduce suboptimal decoders and give a tractable upper bound of error probability using union bound.

Random codebook generation

A simple way is to i.i.d. generate 2^k codewords, and hence the distribution of each codeword $\mathsf{P}(x^n) = \prod_{i=1}^n \mathsf{P}_X(x_i)$. In other words, the elements of the random codebook matrix $\mathbf{C}$ will be i.i.d. distributed according to P_X (P_X is later chosen to be the maximizer in Shannon's capacity formula (1)):

$$\mathbf{C} = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^k) & X_2(2^k) & \cdots & X_n(2^k) \end{bmatrix},$$

and hence

$$\mathsf{P}(\mathbf{c}) \equiv \mathsf{Pr}\left\{\mathbf{C} = \mathbf{c}\right\} = \prod_{w=1}^{2^k} \prod_{i=1}^{n} \mathsf{P}_X(x_i(w)).$$

(we may write $\mathbf{C} \sim \mathsf{P}_X^{\otimes 2^k n}$ for short)

It turns out the symmetry in such codebook ensemble distribution helps simplify the analysis greatly, as we will see later.

Encoding and decoding

For a realization $\mathbf{c}$ of the codebook ensemble $\mathbf{C} \sim \mathsf{P}_X^{\otimes 2^k n}$, we describe the encoding and decoding below:

Encoding: for a message $w \in \{1, ..., 2^k\}$, choose the w-th row of the codebook matrix $\mathbf{c}$ and send it out.

Decoding: ideally one would like to use the following ML decoding rule:

$$\hat{w}_{\mathrm{ML}}(y^n) = \underset{w \in \{1,...,2^k\}}{\arg\max} \ \mathsf{P}(y^n | x^n(w))$$

However, the performance of ML decoder is usually not easy to track, as mentioned before. Instead, we introduce a suboptimal decoder based on robust typical sequences as follows:

$$\boxed{\hat{w}_{\mathrm{T}} = \text{a unique } w \text{ such that } (x^n(w), y^n) \in \mathcal{T}_\varepsilon^{(n)}(X, Y).}$$

Here $(X, Y) \sim \mathsf{P}_X \mathsf{P}_{Y|X}$.

Robust typical sequence

A robust typical sequence is a sequence whose empirical PMF (type) is close to the true per-letter distribution.

For a sequence x^n, its type is given by the frequency of occurrence of a symbol in x^n: $\hat{\mathsf{P}}_{x^n}(a) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i = a\}, \ \forall\, a \in \mathcal{X}$.

Due to LLN, $\hat{\mathsf{P}}_{X^n}(a) \xrightarrow{\text{p}} \mathsf{P}_X(a)$ for all $a \in \mathcal{X}$ as $n \to \infty$, if $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_X$, $i = 1, ..., n$.

Definition 5 (Robust Typical Sequence)

For $\varepsilon \in (0,1)$, a sequence x^n is called ε-robust typical with respect to random variable $X \sim \mathsf{P}_X$ if $\left|\hat{\mathsf{P}}_{x^n}(a) - \mathsf{P}_X(a)\right| \le \varepsilon \mathsf{P}_X(a), \ \forall\, a \in \mathcal{X}$.

The ε-typical set $\boxed{\mathcal{T}_\varepsilon^{(n)}(X) := \{x^n \in \mathcal{X}^n \mid x^n \text{ is } \varepsilon\text{-typical w.r.t. rv } X\}}$.

Equivalent notation: $\mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X)$.

Note: Robust typical sequences also satisfy AEP for memoryless sequences.

Note: When the context is clear, we may write "$\mathcal{T}_\varepsilon^{(n)}$" instead of "$\mathcal{T}_\varepsilon^{(n)}(X)$".

Example 4

Consider a random bit sequence generated i.i.d. based on $\mathrm{Ber}\left(\frac{1}{2}\right)$. Let us set $\varepsilon = 0.2$ and $n = 10$. What is $\mathcal{T}_\varepsilon^{(n)}$? How large is the typical set?

sol: Based on the definition, a n-sequence x^n is ε-typical iff

$$\hat{\mathsf{P}}_{x^n}(0) \in [0.4, 0.6] \text{ and } \hat{\mathsf{P}}_{x^n}(1) \in [0.4, 0.6].$$

In other words, the # of "0"s in the sequence should be 4, 5, or 6. Hence, $\mathcal{T}_\varepsilon^{(n)}$ consists of all length-10 sequences with 4, 5, or 6 "0"s.

The size of $\mathcal{T}_\varepsilon^{(n)}$ is $\binom{10}{4} + \binom{10}{5} + \binom{10}{6} = 714$.

Error probability analysis

Since the ML decoder is optimal, we can analyze the performance of the typicality decoder and use it as an upper bound.

Hence, our goal is turned to showing that if $\mathrm{R} < \mathrm{C}^{\mathrm{I}}$, $\lim_{n\to\infty} \mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right] = 0$.

1 Leverage symmetry in the codebook ensemble to simplify $\mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right]$ and argue that we can focus on the error probability of the first random codeword $X^n(1)$ (its randomness comes from the random $\mathbf{C}$):

$$\begin{aligned}
\mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right] &= \mathsf{E}_{\mathbf{C}}\left[2^{-k}\sum\nolimits_{w=1}^{2^k}\lambda_w(\mathbf{C})\right] = 2^{-k}\sum_{w}\mathsf{E}_{\mathbf{C}}\left[\lambda_w(\mathbf{C})\right] \\
&= 2^{-k}\sum_{w}\mathsf{E}_{\mathbf{C}}\left[\lambda_1(\mathbf{C})\right] = \mathsf{E}_{\mathbf{C}}\left[\lambda_1(\mathbf{C})\right] \\
&= \mathsf{Pr}\left\{\hat{w}_{\mathrm{T}}(Y^n) \neq 1 \,|\, W = 1\right\}
\end{aligned}$$

In the last line, the probability is induced by the random $X^n(1)$ and Y^n.

2 For notational simplicity, use $\mathcal{E}$ to denote the error event $\{\hat{w}_{\mathrm{T}}(Y^n) \neq 1\}$. Our next focus is to upper bound $\Pr\{\mathcal{E} \mid W = 1\} := \mathsf{P}_1(\mathcal{E})$.

The trick here is to distinguish $\mathcal{E}$ into two different kinds of errors:

$$\mathcal{E} = \mathcal{E}_{\mathrm{a}} \cup \mathcal{E}_{\mathrm{t}}, \text{ where } \begin{cases} \mathcal{E}_{\mathrm{a}} := \left\{(X^n(1), Y^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_{X,Y})\right\} \\ \mathcal{E}_{\mathrm{t}} := \left\{(X^n(w), Y^n) \in \mathcal{T}_\varepsilon^{(n)} \text{ for some } w \neq 1\right\} \end{cases}$$

The core is whether or not the pair of sequences $(X^n(w), Y^n)$ are ε-typical with respect to $\mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X}$.

Let us define $\mathcal{A}_w := \left\{(X^n(w), Y^n) \in \mathcal{T}_\varepsilon^{(n)}\right\}$.

We can then rewrite $\mathcal{E}_{\mathrm{a}} = \mathcal{A}_1^c$, $\mathcal{E}_{\mathrm{t}} = \cup_{w\neq 1} \mathcal{A}_w$, and hence

$$\mathcal{E} = \mathcal{E}_{\mathrm{a}} \cup \mathcal{E}_{\mathrm{t}} = \mathcal{A}_1^c \cup \left(\cup_{w\neq 1} \mathcal{A}_w\right).$$

3 We are now ready to apply the union bound:

$$\mathsf{P}_1\left\{\mathcal{E}\right\} = \mathsf{P}_1\left\{\mathcal{A}_1^c \cup \left(\cup_{w\neq 1}\mathcal{A}_w\right)\right\} \le \mathsf{P}_1\left\{\mathcal{A}_1^c\right\} + \sum_{w=2}^{2^k} \mathsf{P}_1\left\{\mathcal{A}_w\right\}.$$

Next, we shall develop upper bounds on

- the probability that the actually sent codeword $X^n(1)$ and the received Y^n are not typical.
- the probability that some other (random) codeword $X^n\ (\neq 1)$ and the received Y^n are typical.

Lemma 2 (A Key Lemma)

$\mathsf{P}_1\left\{\mathcal{A}_1\right\} \ge 1-\varepsilon$ *for* n *large enough, and*

$$\mathsf{P}_1\left\{\mathcal{A}_w\right\} \le 2^{-n(\mathrm{I}(X;Y)-\delta(\varepsilon))}$$

for all $w \neq 1$*, where* $\delta(\varepsilon) \to 0$ *as* $\varepsilon \to 0$*.*

4 Finally, let us put all the above together and apply Lemma 2:

$$\mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right] = \mathsf{Pr}\left\{\mathcal{E} \,\middle|\, W = 1\right\} := \mathsf{P}_1\left\{\mathcal{E}\right\} \le \mathsf{P}_1\left\{\mathcal{A}_1^c\right\} + \textstyle\sum_{w=2}^{2^k} \mathsf{P}_1\left\{\mathcal{A}_w\right\}$$
$$\le \varepsilon + \textstyle\sum_{w=2}^{2^k} 2^{-n(\mathrm{I}(X;Y)-\delta(\varepsilon))} \le \varepsilon + 2^{-n\left(\mathrm{I}(X;Y)-\delta(\varepsilon)-\mathrm{R}-\frac{1}{n}\right)}.$$

As long as $\mathrm{R} < \mathrm{I}(X;Y) - \delta(\varepsilon)$, we are able to make $\mathsf{Pr}\left\{\mathcal{E}\right\} \le 2\varepsilon$ for n large enough, which implies $\lim_{n\to\infty} \mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right] = 0$.

We have shown that as long as $\mathrm{R} < \mathrm{I}(X;Y) - \delta(\varepsilon)$,

$$\lim_{n\to\infty} \mathsf{E}_{\mathbf{C}}\left[\mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{C})\right] = 0,$$

and hence there exists a realization of codebook $\mathbf{c}$ such that $\lim\limits_{n\to\infty} \mathsf{P}_{\mathsf{e},\mathrm{T}}^{(n)}(\mathbf{c}) = 0$.

Finally, taking the codebook generating distribution $\mathsf{P}_X = \arg\max_{\mathsf{P}_X} \mathrm{I}(X;Y)$, we conclude that $\forall\, \mathrm{R} < \mathrm{C}^{\mathrm{I}} = \max_{\mathsf{P}_X} \mathrm{I}(X;Y)$, R is achievable by choosing an appropriate ε. □

Proof of Lemma 2 (1): recap of typicality

Recall: by definition, an ε-typical sequence (x^n, y^n) shall satisfy

$$\left|\hat{\mathsf{P}}_{x^n,y^n}(a,b) - \mathsf{P}_{X,Y}(a,b)\right| \le \varepsilon\,\mathsf{P}_{X,Y}(a,b),\ \forall\,(a,b) \in \mathcal{X} \times \mathcal{Y}.$$

(Note: we can think of (X, Y) as a r.v. and apply the same definition of typicality!)

Hence, if $(X^n, Y^n) \sim \mathsf{P}_{X,Y}^{\otimes n}$, then we have

0 $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_{X,Y}) \implies x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X),\ y^n \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_Y)$.

1 $\forall\,(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}$, $\left|\frac{1}{n}\log \mathsf{P}(x^n, y^n) + \mathrm{H}(X,Y)\right| \le \delta(\varepsilon)$, $\delta(\varepsilon) = \varepsilon \mathrm{H}(X,Y)$.

2 $\mathsf{P}(\mathcal{T}_\varepsilon^{(n)}) \ge 1 - \varepsilon$ for n large enough.

3 $|\mathcal{T}_\varepsilon^{(n)}| \le 2^{n(\mathrm{H}(X,Y)+\delta(\varepsilon))}$.

4 $|\mathcal{T}_\varepsilon^{(n)}| \ge (1-\varepsilon)2^{n(\mathrm{H}(X,Y)-\delta(\varepsilon))}$ for n large enough.

Proof of Lemma 2 (2): typical with the true codeword

Let us first consider $\mathsf{P}_1\left\{\mathcal{A}_1\right\} = \mathsf{Pr}\left\{(X^n(1), Y^n) \in \mathcal{T}_\varepsilon^{(n)} \middle| W = 1\right\}$.

We are averaging over a random codebook ensemble $\mathbf{C}$, and the random codebook is generated element-by-element i.i.d. based on P_X.

DMC without feedback implies $\mathsf{P}(y^n|x^n) = \prod_{i=1}^n \mathsf{P}_{Y|X}(y_i|x_i)$.

Hence, given $W = 1$, $(X^n(1), Y^n)$ has the following joint distribution:

$$\begin{aligned}\mathsf{P}(x^n, y^n) &= \mathsf{P}(x^n) \cdot \mathsf{P}(y^n|x^n) = \textstyle\prod_{i=1}^n \mathsf{P}_X(x_i) \cdot \prod_{i=1}^n \mathsf{P}_{Y|X}(y_i|x_i) \\ &= \textstyle\prod_{i=1}^n \mathsf{P}_{X,Y}(x_i, y_i)\end{aligned}$$

By Property 2 (LLN) in the previous slide, we see that for n large enough,

$$\mathsf{P}_1\left\{\mathcal{A}_1\right\} = \mathsf{Pr}\left\{(X^n(1), Y^n) \in \mathcal{T}_\varepsilon^{(n)} \,\middle|\, W = 1\right\} \geq 1 - \varepsilon.$$

Proof of Lemma 2 (3): typical with a wrong codeword

Consider $\mathsf{P}_1\left\{\mathcal{A}_w\right\} = \mathsf{Pr}\left\{(X^n(w), Y^n) \in \mathcal{T}_\varepsilon^{(n)} \,\middle|\, W = 1\right\}$ for $w \neq 1$.

Note that we are averaging over a random codebook ensemble $\mathbf{C}$, and the random codebook is generated element-by-element i.i.d. based on P_X.

$X^n(1)$ and $X^n(w)$ have the same distribution $\mathsf{P}_X^{\otimes n}$ but they are *independent*.

Due to DMC, $(X^n(1), Y^n) \perp\!\!\!\perp X^n(w)$. Hence, $Y^n \perp\!\!\!\perp X^n(w)$, and

$$\begin{aligned} \mathsf{P}_1\left\{\mathcal{A}_w\right\} &= \textstyle\sum_{(x^n,y^n)\in\mathcal{T}_\varepsilon^{(n)}} \mathsf{P}(x^n)\cdot\mathsf{P}(y^n) \\ &\le \underbrace{2^{n(1+\varepsilon)\mathrm{H}(X,Y)}}_{\substack{\text{cardinality upper}\\ \text{bound on typical set}}} \cdot \underbrace{2^{-n(1-\varepsilon)\mathrm{H}(X)}}_{\substack{\text{upper bound on prob.}\\ \text{of a typical sequence}}} \cdot \underbrace{2^{-n(1-\varepsilon)\mathrm{H}(Y)}}_{\substack{\text{upper bound on prob.}\\ \text{of a typical sequence}}} \\ &= 2^{-n(\mathrm{I}(X;Y)-\delta(\varepsilon))}, \end{aligned}$$

where $\delta(\varepsilon) = \varepsilon\left(\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y)\right) \to 0$ as $\varepsilon \to 0$. □

Some reflections

Reflection 1: **Mutual independence of codewords**

In the random coding argument of the proof, $2^k n$ elements of the codebook matrix $\mathbf{C}$ are generated i.i.d., and hence the 2^k rows $\left\{X^n(1), ..., X^n(2^k)\right\}$ are *mutually independent*. In the proof, only pairwise independence is required: $X^n(1) \perp\!\!\!\perp X^n(w),\ \forall\, w \neq 1$.

Reflection 2: **Typicality decoder**

We use typicality decoder other than the optimal ML decoder to find tractable upper bounds on the error probability. There are other suboptimal decoders can be used. For example, the following threshold decoder can also work:

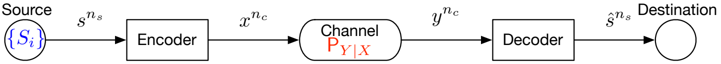
$$\boxed{\hat{w}_{\text{th}} := \text{a unique } w \text{ such that } \iota\left(x^n(w); y^n\right) > \beta},$$

where $\iota\left(x^n; y^n\right) := \log \frac{\mathsf{P}(x^n, y^n)}{\mathsf{P}(x^n)\mathsf{P}(y^n)} = \sum_{k=1}^n \log \frac{\mathsf{P}_{Y|X}(y_k|x_k)}{\mathsf{P}_X(x_k)}$, called the *information density function*, and $\beta := n(\mathrm{I}(X;Y) - \varepsilon)$.

1. Mutual Information
   - Definitions
   - Properties

2. Noisy Channel Coding Theorem
   - Channel Capacity
   - Proof of Weak Converse
   - Feedback Capacity
   - Proof of Achievability
   - Source-Channel Separation

Joint source-channel coding: problem setup



Source model: DMS P_S with entropy $\mathrm{H}(\mathsf{P}_S)$.

Channel model: DMC $\mathsf{P}_{Y|X}$ with channel capacity $\mathrm{C}(\mathsf{P}_{Y|X})$.

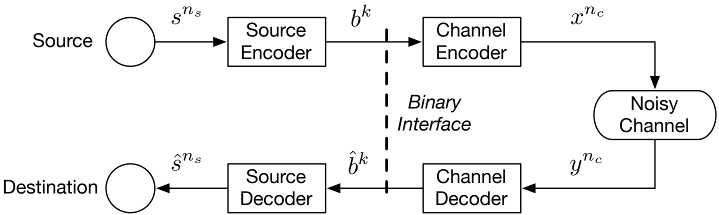
1. A (n_s, n_c) joint source-channel code consists of
   - $\mathrm{enc}_{n_c} : \mathcal{S}^{n_s} \to \mathcal{X}^{n_c}$ that maps source s^{n_s} to a codeword x^{n_c}.
   - $\mathrm{dec}_{n_c} : \mathcal{Y}^{n_c} \to \mathcal{S}^{n_s}$ that maps a channel output y^{n_c} to a reconstructed $\hat{s}^{n_s}$.
2. The error probability is defined as $\mathsf{P}_{\mathsf{e}}^{(n_c)} := \Pr\{S^{n_s} \neq \hat{S}^{n_s}\}$.
3. R is achievable if there exist a sequence of $(\lceil n_c \mathrm{R} \rceil, n_c)$ codes such that $\lim_{n_c \to \infty} \mathsf{P}_{\mathsf{e}}^{(n_c)} = 0$.

Source-channel separation theorem

Theorem 6 (Source-Channel Separation)

If $\mathrm{R} < \mathrm{C}(\mathsf{P}_{Y|X})/\mathrm{H}(\mathsf{P}_S)$*, then* R *is achievable, i.e., lossless reconstruction of source* $\{S_i\}$ *is possible.*

Conversely, if $\mathrm{R} > \mathrm{C}(\mathsf{P}_{Y|X})/\mathrm{H}(\mathsf{P}_S)$*, then* R *is not achievable, i.e., lossless reconstruction is impossible.*



pf: (Achievability Part):

- Choose an $(n_s, \lfloor n_s \mathrm{R}_s \rfloor)$ lossless source code with $\mathrm{R}_s = \mathrm{H}(\mathsf{P}_S) + \varepsilon_s$.
- Choose an $(n_c, \lceil n_c \mathrm{R}_c \rceil)$ channel code with $\mathrm{R}_c = \mathrm{C}(\mathsf{P}_{Y|X}) - \varepsilon_c$.

By the channel coding theorem, the binary sequence b^k lying at the digital interface between the source and the channel coders can be decoded with vanishing error probability.

By the lossless source coding theorem, the source sequences can be reconstructed with vanishing error probability as long as b^k can be successfully decoded by the channel decoder.

Concatenate the above two codes together, we see that if

$$n_s \mathrm{R}_s < n_c \mathrm{R}_c \iff \tfrac{n_s}{n_c} < \tfrac{\mathrm{R}_c}{\mathrm{R}_s} = \tfrac{\mathrm{C} - \varepsilon_c}{\mathrm{H}(\mathsf{P}_S) + \varepsilon_s},$$

the separation scheme is able to reconstruct the source sequence with vanishing error probability. Since $\varepsilon_s, \varepsilon_c$ can be made arbitrarily small, as long as $\mathrm{R} < \frac{\mathrm{C}}{\mathrm{H}(\mathsf{P}_S)}$, it is achievable. □

(Converse Part):

We shall prove that for all achievable R, $\mathrm{R} \le \frac{\mathrm{C}}{\mathrm{H}(\mathsf{P}_S)}$.

$$
\begin{aligned}
& n_s \mathrm{H}(\mathsf{P}_S) \\
& = \mathrm{H}(S^{n_s}) = \mathrm{I}\left(S^{n_s}; \hat{S}^{n_s}\right) + \mathrm{H}\left(S^{n_s} \middle| \hat{S}^{n_s}\right) && (5) \\
& \le \mathrm{I}(S^{n_s}; Y^{n_c}) + \left(1 + \mathsf{P}_{\mathsf{e}}^{(n_c)} n_s \log|\mathcal{S}|\right) && (6) \\
& \le \sum_{k=1}^{n_c} \mathrm{I}\left(S^{n_s}; Y_k \middle| Y^{k-1}\right) + \left(1 + \mathsf{P}_{\mathsf{e}}^{(n_c)} n_s \log|\mathcal{S}|\right) \le n_c\left(\mathrm{C} + \varepsilon_{n_c}\right), && (7)
\end{aligned}
$$

- (5) is due to the property of DMS and chain rule.
- (6) is due to $S^{n_s} - Y^{n_c} - \hat{S}^{n_s}$ and Fano's inequality.
- (7) is due to similar steps as in the channel coding converse proof.

Since $\varepsilon_{n_c} \to 0$ as $n_c \to \infty$, $\mathrm{R} \le \frac{n_s}{n_c} \le \frac{\mathrm{C}}{\mathrm{H}(\mathsf{P}_S)}$ if R is achievable. □

Summary: noisy channel coding and capacity

- For DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$ with or without feedback: $\mathrm{C} = \max_{\mathsf{P}_X} \mathrm{I}(X;Y)$.
- Symmetric channel capacity $= \log|\mathcal{Y}| - \mathrm{H}(\boldsymbol{p})$, where $\boldsymbol{p}$ permutes all rows of $\mathsf{P}_{Y|X}$. Erasure channel capacity $= (1-p)\log|\mathcal{X}|$.
- Blahut-Arimoto algorithm for computing channel capacity.
- Weak converse: Fano's inequality, data processing, DMC assumption.
- Achievability: random coding argument, typicality decoder.
- Feedback does not increase the capacity of DMC.
- Joint source-channel coding theorem:

$$\mathrm{R} < \tfrac{\mathrm{C}}{\mathrm{H}(\mathsf{P}_S)} \implies \mathrm{R} \text{ is achievable}$$

$$\mathrm{R} > \tfrac{\mathrm{C}}{\mathrm{H}(\mathsf{P}_S)} \implies \mathrm{R} \text{ is not achievable}$$

- Source-channel separation is optimal.