LECTURE 6
COMMUNICATION OVER A NOISY CHANNEL

Information Theory

Thomas Debris-Alazard

Inria, École Polytechnique

MOTIVATION

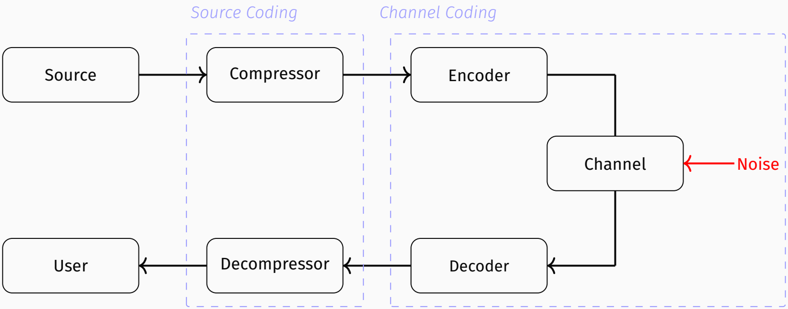
$$\textit{Up to now: source coding } \left(\textit{compression}\right) \textit{ with } \underbrace{\textit{block codes}}_{\textit{Shannon th.}}, \underbrace{\textit{symbol codes}}_{\textit{Huffman}} \textit{ and } \underbrace{\textit{stream codes}}_{\textit{Arith Coding}}$$

Implicitly: channel from the compressor to the decompressor was noise-free. . .

Issues:

- Channels in real life are noisy. . .
- If our aim is to transmit information, not to compress: how can we reliably transmit information over a noisy channel?

THE BIG PICTURE



A FALSE INTUITION

Suppose we transmit 1000 bits per second with $p_0 = p_1 = 1/2$ over a noisy channel that flips bits with probability $f = 0.1$. What is the rate of transmission of information?

It is false to guess 900 bits per second. . .

$\longrightarrow$ We don't know **where** the errors occurred!

WHAT WE HAVE LEARNT UP TO NOW

Reasonable thought:

a measure of the information transmitted across a noisy channel is given by the

entropy of the source minus the conditional entropy of the source given the received signal

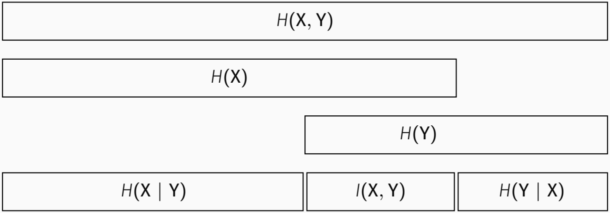
$\longrightarrow$ It is the mutual information!

AN IMPORTANT REMINDER: MUTUAL INFORMATION

$$H(\mathrm{X}) = \sum_x p(x) \log_2 1/p(x) \quad \text{and} \quad H(\mathrm{X} \mid \mathrm{Y}) = \sum_{x,y} p(x,y) \log_2 1/(p(x \mid y))$$

$$I(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X} \mid \mathrm{Y}) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- $I(\mathrm{X},\mathrm{Y}) = I(\mathrm{Y},\mathrm{X}) = H(\mathrm{Y}) - H(\mathrm{Y} \mid \mathrm{X})$
- $I(\mathrm{X},\mathrm{Y}) \geq 0$
- $I(\mathrm{X},\mathrm{Y}) = D_{\mathrm{KL}}(p(x,y)||p(x)p(y))$



$H(\mathrm{X},\mathrm{Y})$

$H(\mathrm{X})$

$H(\mathrm{Y})$

$H(\mathrm{X} \mid \mathrm{Y})$ $I(\mathrm{X},\mathrm{Y})$ $H(\mathrm{Y} \mid \mathrm{X})$

COURSE OUTLINE

1. Noisy Channels and Capacity
2. Noisy-Channel Coding Theorem
3. Proof of Achievability
4. What is Impossible
5. Conclusion

A WARNING

Be careful, this lecture is "mathematically" very abstract

$\longrightarrow$ Lecture 7: we prove Shannon's second theorem in a particular case

$\big($linear codes and Hamming metric$\big)$

where the geometric perspective works especially well, making the proofs much more intuitive

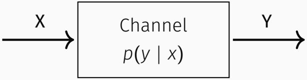
NOISY CHANNELS AND CAPACITY

DISCRETE CHANNEL

- $\mathcal{X}$ source alphabet $\left(X \in \mathcal{X}\right)$
- $\mathcal{Y}$ output alphabet after transmitting source symbols across a noisy channel $\left(Y \in \mathcal{Y}\right)$

Discrete Channel:

System consisting of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a set of probability distributions matrix $p(y \mid x)$



⟶ Restriction: **memoryless channels**, the probability of observing y depends only on the source symbol x transmitted

$\left(\text{in particular: the } i\text{thm output } y_i \text{ only depends of } x_i \text{ and not previous sent } x_1, \ldots, x_{i-1}\right)$

Discrete Memoryless Channel Q:

Characterized by an input/source alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and a set of probability distributions matrix $p(y \mid x)$. The probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs, *i.e.*,

$$\forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^N \times \mathcal{Y}^n, \;\; p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^N p(y_i \mid x_i)$$

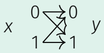
Transition probabilities can be written as matrix

$$Q_{i,j} = p(y = b_i \mid x =_j)$$

Each column of $\mathbf{Q}$ is a probability vector

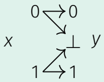
SOME EXAMPLES

- Binary Symmetric Channel: $\mathcal{X} = \mathcal{Y} = \{0, 1\}$



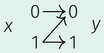
$$p(y=0 \mid x=0) = 1-f \qquad p(y=0 \mid x=1) = f$$
$$p(y=1 \mid x=0) = f \qquad p(y=1 \mid x=1) = 1-f$$

- Binary Erasure Channel: $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, \perp\}$



$$p(y=0 \mid x=0) = 1-f \qquad p(y=0 \mid x=1) = 0$$
$$p(y=\perp \mid x=0) = f \qquad p(y=\perp \mid x=1) = f$$
$$p(y=1 \mid x=0) = 0 \qquad p(y=1 \mid x=1) = 1-f$$

- Z-channel: $\mathcal{X} = \mathcal{Y} = \{0, 1\}$



$$p(y=0 \mid x=0) = 1 \qquad p(y=0 \mid x=1) = f$$
$$p(y=1 \mid x=0) = 0 \qquad p(y=1 \mid x=1) = 1-f$$

- Noisy typewriter: $\mathcal{X} = \mathcal{Y} = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}, -\}$

$$p(- \mid \mathrm{A}) = 1/3 \qquad p(\mathrm{A} \mid \mathrm{B}) = 1/3$$
$$p(\mathrm{A} \mid \mathrm{A}) = 1/3 \qquad p(\mathrm{B} \mid \mathrm{B}) = 1/3 \qquad \cdots$$
$$p(\mathrm{B} \mid \mathrm{A}) = 1/3 \qquad p(\mathrm{C} \mid \mathrm{B}) = 1/3$$

INFERRING THE INPUT GIVEN THE OUTPUT

From source x to output y: joint ensemble **XY**:

$$p(x, y) = p(y \mid x) \cdot p(x)$$

If we receive y, what was the input symbol x?

Posterior distribution:

$$p(x \mid y) = \frac{p(y \mid x) \cdot p(x)}{p(y)} = \frac{p(y \mid x) \cdot p(x)}{\sum_{x'} p(y \mid x') \cdot p(x')}$$

SOME EXAMPLES

Binary Symmetric Channel:

Consider a probability of error $f = 0.15$ and source $\mathbf{X} = (p(0) = 0.9, p(1) = 0.1)$. Assume we observe 1,

$$\begin{aligned} p(x = 1 \mid y = 1) &= \frac{p(y = 1 \mid x = 1)p(x = 1)}{\sum_{x' \in \{0,1\}} p(y \mid x')p(x')} \\ &= \frac{0.85 \cdot 0.1}{0.85 \cdot 0.1 + 0.15 \cdot 0.9} \\ &= 0.39 \end{aligned}$$

The “$x = 1$” is still less probable than “$x = 0$” although it is not as improbable as before

Z Channel:

Consider a probability of error $f = 0.15$ and source $\mathbf{X} = (p(0) = 0.9, p(1) = 0.1)$. Assume we observe 1,

$$\begin{aligned} p(x = 1 \mid y = 1) &= \frac{0.85 \cdot 0.1}{0.85 \cdot 0.1 + 0 \cdot 0.9} \\ &= 1 \end{aligned}$$

Given the output “$y = 1$”, we become certain of the input

INFORMATION CONVEYED BY A CHANNEL: MUTUAL INFORMATION

How much information can be communicate through a noisy channel?

Information meaning: rate of transmission

$$\frac{\text{number of recovered bits from } N \text{ received bits}}{N}$$

$\longrightarrow$ We want to find ways of using the channel such that we recover the bits which were wanted to be communicated with negligible probability of error!

How much information the output Y conveys about the input X:

$$I(X, Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- We think $I(X, Y)$ as $H(X) - H(X \mid Y)$
- For computations, often easier to evaluate $H(Y) - H(Y \mid X)$ which is equal to $I(X, Y)$ $\Big($*by definition of the channel we know the* $p(y \mid x)$*'s*$\Big)$

SOME EXAMPLES

Binary Symmetric Channel:

Consider a probability of error $f = 0.15$ and source $\mathbf{X} = (p(0) = 0.9, p(1) = 0.1)$. We have

$$p(y = 0) = 0.78 \quad \text{and} \quad p(y = 1) = 0.22$$

Then,

$$I(\mathbf{X}, \mathbf{Y}) = 0.15$$

We can communicate 15 bits by sending 100bits

***Z*-channel:**

Consider a probability of error $f = 0.15$ and source $\mathbf{X} = (p(0) = 0.9, p(1) = 0.1)$. We have

$$p(y = 0) = 0.915 \quad \text{and} \quad p(y = 1) = 0.085$$

Then,

$$I(\mathbf{X}, \mathbf{Y}) = 0.36$$

We can communicate 36 bits by sending 100bits

$\longrightarrow$ Given $\mathbf{X}$, the Z-channel is more reliable!

THE CAPACITY

Mutual information between input and output depends on the chosen input X

Furthermore, the output Y is only function of $p(x)$ and the channel transition distribution $p(y \mid x)$

$$\left(p(y) = \sum_x p(y \mid x) p(x)\right)$$

$\longrightarrow$ We want to maximize the mutual information by choosing the best input distribution X

Capacity:

Given a channel Q, its capacity is:

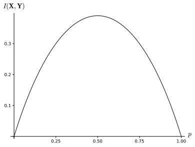
$$C(Q) \overset{\text{def}}{=} \max_{\mathrm{X}} I(\mathrm{X}, \mathrm{Y})$$

SOME EXAMPLES

Binary Symmetric Channel:

Given the binary symmetric channel with probability of error $f = 0.15$ and $\mathbf{X} = (p(0) = 1 - p, p(1) = p)$,

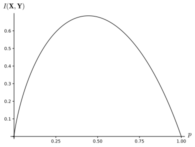
$$I(\mathbf{X}, \mathbf{Y}) = h((1-f)p + f(1-p)) - h(f) \quad \text{and} \quad C(Q) = 1 - h(f)$$



Z-channel:

Given the Z channel with probability of error $f = 0.15$ and $\mathbf{X} = (p(0) = 1 - p, p(1) = p)$,

$$I(\mathbf{X}, \mathbf{Y}) = h(p(1-f)) - ph(f)$$



BLOCK CODE

(N, K)-block code:

For a channel Q, an (N, K)-block code is a list of 2^K codewords:

$$\mathcal{C} = \left\{\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \dots, \mathrm{x}^{(2^K)}\right\}, \quad \mathrm{x}^{(s)} \in \mathcal{X}^N$$

$\longrightarrow$ In the literature block codes are also called "*codes*" or "*non-linear codes*"

Using this code $\mathcal{C}$:

- **Encoding:** $s \in \left\{1, 2, \dots, 2^K\right\}$ encoded as $\mathrm{x}^{(s)} \in \mathcal{X}^N$
- **Decoding:** a map from $\mathcal{Y}^N$ to $\widehat{s} \in \{0, 1, \dots, 2^K\}$ where the extra symbol 0 is "failure"

Rate:

The rate of an (N, K)-block code is

$$R \overset{\text{def}}{=} \tfrac{K}{N} \in [0, \log_2 \sharp\mathcal{X}]$$

$\longrightarrow$ To transmit K bits, we send codewords having $N \log_2 \sharp\mathcal{X} \geq K$ bits

$\left(N \log_2 \sharp\mathcal{X} - K \text{ is the number of redundancy bits}\right)$

PROBABILITY OF BLOCK ERROR

Given a channel, a distribution over the signal to encode $p(s_{\text{in}})$, a decoder may fail to recover s_{in}.

We distinguish two probabilities of error:

- The probability of block error $\Big($average$\Big)$:

$$p_{\text{B}} \overset{\text{def}}{=} \sum_{s_{\text{in}}} p(s_{\text{out}} \neq s_{\text{in}} \mid s_{\text{in}}) p(s_{\text{in}})$$

- The maximal probability of block error $\Big($worst-case$\Big)$:

$$p_{\text{BM}} \overset{\text{def}}{=} \max_{s_{\text{in}}} p(s_{\text{out}} \neq s_{\text{in}} \mid s_{\text{in}}) \quad \Big(\text{it does not depend on the } s_{\text{in}} \text{ distribution}\Big)$$

If $p_{\text{BM}} \leq \varepsilon$, then $p_{\text{B}} \leq \varepsilon$

- The optimal decoder: it minimizes p_{BM}, *i.e.*, given $\mathbf{y} \in \mathcal{Y}^N$ it outputs $\widehat{s}_{\text{opt}}$,

$$\widehat{s}_{\text{opt}} = \arg\max_{s} p(s \mid \mathbf{y}) \quad \text{where} \quad p(s \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid s)p(s)}{\sum_{s'} p(y \mid s')p(s')}$$

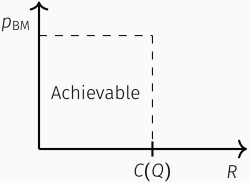
A particular case: maximum likelihood decoder

If the s_{in} are supposed to be uniformly distributed in $\{1, \ldots, 2^K\}$, then optimal decoder is maximum likelihood decoder,

$$\widehat{s}_{\text{opt}} = \arg\max p(\mathbf{y} \mid s) \quad \Big(\text{we inverted } s \text{ and } \mathbf{y}\Big)$$

Shannon's noisy-channel coding theorem $\big($positive part...$\big)$

Given a discrete memoryless channel Q, for any $\varepsilon > 0$ and $R < C(Q)$: for large enough N, there exists a block code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error p_{BM} is $< \varepsilon$



It proves that we can reliably transmit information but it has a cost, the rate cannot a priori be 1

$\longrightarrow$ But good news: the rate is > 0

AN EXAMPLE

- Noisy typewriter: $\mathcal{X} = \mathcal{Y} = \{\mathtt{A}, \mathtt{B}, \ldots, \mathtt{Z}, -\}$

$$\begin{array}{rlc} p(- \mid \mathtt{A}) = 1/3 & p(\mathtt{A} \mid \mathtt{B}) = 1/3 & \\ p(\mathtt{A} \mid \mathtt{A}) = 1/3 & p(\mathtt{B} \mid \mathtt{B}) = 1/3 & \cdots \\ p(\mathtt{B} \mid \mathtt{A}) = 1/3 & p(\mathtt{C} \mid \mathtt{B}) = 1/3 & \end{array}$$

A clever code: use the $(1, \log_2 9)$-block code consisting of the 9 letters $\mathtt{B}, \mathtt{E}, \mathtt{H}, \ldots, \mathtt{Z}$

These letters form a non-confusable subset of the input alphabet!

$\longrightarrow$ Decoding is easy and will succeed with probability one $\big((\mathtt{A}, \mathtt{B}, \mathtt{C}) \mapsto \mathtt{B}, (\mathtt{D}, \mathtt{E}, \mathtt{F}) \mapsto \mathtt{E}, \text{etc.}\ldots\big)$

What a surprise:

The capacity of the noisy typewriter is $\log_2 9$

AN INTUITIVE STATEMENT

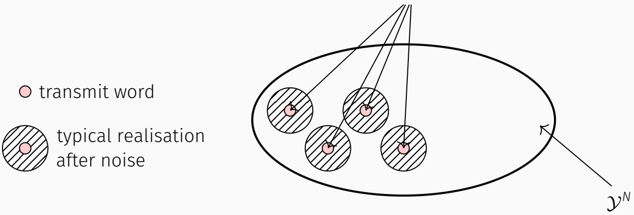
The theorem	How it applies to the noisy typewriter
For any discrete memoryless channel Q,	$C(Q) = \log_2 9$
for all $\varepsilon > 0$ and $R < C(Q)$,	*no matter what $\varepsilon > 0$ and $R < \log_2 9$ are*
for large enough N,	*we set the block length $N = 1$*
there exists a block code of length N	*The block-code is $\{\mathtt{B}, \mathtt{E}, \ldots, \mathtt{Z}\}$. The value K*
and rate $\geq R$	*is given by $2^K = 9$, so $K = \log_2 9$. This code has rate*
	$\log_2 9$ which is greater than the requested value R
and a decoding algorithm	*The decoding maps the received letter to the*
	nearest letter in the code
such that the maximal probability of	*the maximal probability of block error is zero*
block error is $< \varepsilon$	*which is $< \varepsilon$*

THE BIG PICTURE: AVOID CONFUSION

What we learn with the noisy typewriter:

Use a code for which there are no confusions after sending through the noisy channel. . .

$2^{NC(Q)}$ words can be transmitted without confusion



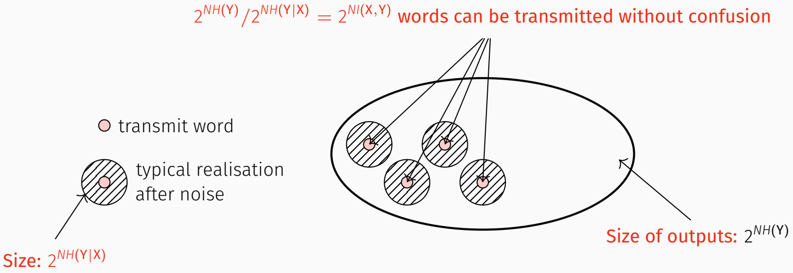
○ transmit word

typical realisation after noise

$\mathcal{Y}^N$

AN INTUITIVE PROOF

1. Given a sequence $\mathbf{x} \in \mathcal{X}^N$ drawing from X, there are $2^{NH(\mathsf{Y}|\mathsf{X})}$ possible outputs
2. The size of possible outputs is $2^{NH(\mathsf{Y})}$
3. To avoid confusion, we can transmit at most $2^{NH(\mathsf{Y})}/2^{NH(\mathsf{Y}|\mathsf{X})} = 2^{NI(\mathsf{X},\mathsf{Y})}$ codewords
4. Then we choose X to maximize the number of possible codewords to transmit


$2^{NH(\mathsf{Y})}/2^{NH(\mathsf{Y}|\mathsf{X})} = 2^{NI(\mathsf{X},\mathsf{Y})}$ words can be transmitted without confusion
transmit word
typical realisation after noise
Size: $2^{NH(\mathsf{Y}|\mathsf{X})}$
Size of outputs: $2^{NH(\mathsf{Y})}$

NOISY-CHANNEL CODING

THE THEOREM

Shannon's noisy-channel coding theorem has two parts: one positive and one negative

Shannon's noisy-channel coding theorem:

1. For every discrete memoryless channel Q, the channel capacity

   $$C(Q) \overset{\text{def}}{=} \max_{\mathsf{X}} I(\mathsf{X}, \mathsf{Y})$$

   has the following property: for all $\varepsilon > 0$ and $R < C(Q)$, for large enough N, there exists a block code of length N and rate $\geq R$, and decoding algorithm such that the maximal probability of block error p_{BM} is $< \varepsilon$

2. Reciprocally, given a of sequence (N, RN)-codes, if the maximal probability of block error is tending to 0 $\big(\text{with } N\big)$, then necessarily $R \leq C(Q)$

OUR AIM

To prove this theorem!

The key tool:

Given a code $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(2^K)}\right\}$, we will decode $\mathbf{y}$ as s if $\left(\mathbf{x}^{(s)}, \mathbf{y}\right)$ are jointly typical

The proof will centre on determining the probabilities that

- the true input is not jointly typical with the received/output sequence
- a false input/codeword is jointly typical with the received/output sequence

$\longrightarrow$ We will show that both probabilities are tending to 0 as soon as the number of codewords is smaller than $2^{NC(Q)}$ and $\mathbf{X}$ being the input distribution maximizing $I(\mathbf{X}, \mathbf{Y})$

JOINTLY TYPICAL SEQUENCES

$\Big($Remember the AEP from Lecture 3$\Big)$

Joint typicality:

A pair of sequence $\mathbf{x}, \mathbf{y}$ of length N are defined to be jointly typical to tolerance β with respect to $p(\mathbf{x}, \mathbf{y})$ $\Big($where $p(\mathbf{x})$ and $p(\mathbf{y})$ are the marginal distributions of $\mathbf{x}$ and $\mathbf{y}\Big)$ if

- $\mathbf{x}$ is typical of $p(\mathbf{x})$, *i.e.*, $\left|\frac{1}{N}\log_2\frac{1}{p(\mathbf{x})} - H(X)\right| < \beta$
- $\mathbf{y}$ is typical of $p(\mathbf{y})$, *i.e.*, $\left|\frac{1}{N}\log_2\frac{1}{p(\mathbf{y})} - H(Y)\right| < \beta$
- $\mathbf{x}, \mathbf{y}$ is typical of $p(\mathbf{x}, \mathbf{y})$, *i.e.*, $\left|\frac{1}{N}\log_2\frac{1}{p(\mathbf{x}, \mathbf{y})} - H(X, Y)\right| < \beta$

The jointly typical set $J_{N\beta}$ is the set of all jointly typical sequence to tolerance β with length N

JOINT TYPICALITY THEOREM

Joint typicality theorem $\left(\text{also called joint AEP theorem}\right)$:

Let $\mathbf{x}, \mathbf{y}$ be picked according to $(XY)^{\otimes N}$, *i.e.*,

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} p(x_i, y_i)$$

1. The probability that $\mathbf{x}, \mathbf{y}$ are jointly typical $\left(\text{to tolerance } \beta\right)$ tends to 1 as $N \to +\infty$
2. $$\sharp J_{N\beta} \leq 2^{N(H(X,Y)+\beta)}$$
3. $\mathbf{x}'$ and $\mathbf{y}'$ are independently distributed according to $p(\mathbf{x})$ and $p(\mathbf{y})$, then
$$p\left((\mathbf{x}', \mathbf{y}') \in J_{N\beta}\right) \leq 2^{-N(I(X,Y)+3\beta)}$$

In Lecture 5 we have proven thanks to Sanov's theorem a stronger result:

$$p\left((\mathbf{x}', \mathbf{y}') \in J_{N\beta}\right) \overset{\text{(poly)}}{=} 2^{-NI(X,Y)}$$

THE PROOF

Proof:

1. Consequence of the weak law of large number

2. $$1 = \sum p(\mathrm{x}, \mathrm{y}) \geq \sum_{J_{N\beta}} p(\mathrm{x}, \mathrm{y}) \geq \sharp J_{N\beta}\ 2^{-N(H(\mathrm{X},\mathrm{Y})+\beta)}$$

3. $$\begin{aligned} p\left((\mathrm{x}', \mathrm{y}') \in J_{N\beta}\right) &= \sum_{(\mathrm{x},\mathrm{y}) \in J_{N\beta}} p(\mathrm{x}) p(\mathrm{y}) \\ &\leq \sharp J_{N\beta}\ 2^{-N(H(\mathrm{X})-\beta)}\ 2^{-N(H(\mathrm{Y})-\beta)} \\ &\leq 2^{N(H(\mathrm{X},\mathrm{Y})+\beta)}\ 2^{-N(H(\mathrm{X})+H(\mathrm{Y})-2\beta)} \\ &= 2^{-N(I(\mathrm{X},\mathrm{Y})+3\beta)} \end{aligned}$$

   where in the last line we used $\left(\text{see Slide } 5\right)$:

   $$H(\mathrm{X}, \mathrm{Y}) - H(\mathrm{X}) - H(\mathrm{Y}) = H(\mathrm{Y} \mid \mathrm{Y}) - H(\mathrm{Y}) = -I(\mathrm{X}, \mathrm{Y})$$

AN INTUITIVE PROOF

By the AEP

- The number of independent pairs $(\mathsf{x}', \mathsf{y}')$ is $2^{NH(\mathsf{X})}2^{NH(\mathsf{Y})} = 2^{N(H(\mathsf{X})+H(\mathsf{Y}))}$
- The number of jointly typical pair $(\mathbf{x}, \mathbf{y})$ is $2^{NH(\mathsf{X},\mathsf{Y})}$
- The probability of hitting $\Big($for independent drawing$\Big)$ a jointly typical pair is roughly given by

$$\frac{2^{NH(\mathsf{X},\mathsf{Y})}}{2^{N(H(\mathsf{X})+H(\mathsf{Y}))}} = 2^{-NI(\mathsf{X},\mathsf{Y})}$$

$\Big($think as "we are always typical"$\Big)$

$\longrightarrow$ Therefore if the number of codewords is $2^{NI(\mathsf{X},\mathsf{Y})}$ we expect that if $\mathbf{x}^{(s)}$ is sent and $\mathbf{y}$ received,

there won't be $x^{(s')}$ with $s' \neq s$ jointly typical with $\mathbf{y}$

NOISY-CHANNEL CODING: ACHIEVABILITY PROOF

ACHIEVABILITY

We want to show that for any rate $R < C(Q)$, there exists a code with this rate such that its maximal probability of block error p_{BM} when decoding tends to 0 when $N \to +\infty$

We consider the following encoding-decoding system with rate R':

1. We fix $p(x)$ and generate the (N, NR')-block code $\mathcal{C}$ consisting of random $\mathbf{x}$ picked as
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i)$$
2. The code is known to both the sender and receiver $\Big($*non-efficient, need to store exponential tables...*$\Big)$
3. For a message $s \in \{1, 2, \ldots, 2^{NR'}\}$ we transmit $\mathbf{x}^{(s)}$. The received signal is $\mathbf{y}$ with,
$$p(\mathbf{y} \mid \mathbf{x}^{(s)}) = \prod_{i=1}^{N} p(y_i \mid x_i^{(s)})$$
4. The signal is decoded by typical set decoding $\Big($*non-efficient, exponential number of probabilities to compute...*$\Big)$

$\Big($the breakthrough idea of Shannon lies in 1., choose a random code!$\Big)$

TYPICAL SET DECODING

Typical Set Decoding:

Given $\mathbf{y}$, we decode it as $\widehat{s}$ if

1. $\left(\mathbf{x}^{(\widehat{s})}, \mathbf{y}\right)$ are jointly typical and,
2. there is no other s' such that $\left(\mathbf{x}^{(s')}, \mathbf{y}\right)$ are jointly typical

Be careful:

This procedure is not efficient, we need

1. to store all the $2^{NR'}$-codewords. . .
2. to compute $2^{NR'}$ probabilities to decide that a pair is jointly typical or not. . .

Our goal: to analyze the maximal probability of errors of the typical set decoding

PROBABILITIES OF ERRORS

First, our choice of the code $\mathcal{C}$ is random

$$\text{Given } \mathcal{C} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(2^{NR'})}\right), \quad p(\mathcal{C}) = \prod_{s=1}^{2^{NR'}} \prod_{i=1}^{N} p(x_i^{(s)}) = \prod_{s=1}^{2^{NR'}} p(\mathbf{x}^{(s)})$$

During the decoding there are three probabilities of errors that interest us:

- Probability of block error for a fixed code $\mathcal{C}$ $\Big($be careful the probability is also over s$\Big)$,

$$p_{\mathrm{B}}(\mathcal{C}) \overset{\text{def}}{=} p(\widehat{s} \neq s \mid \mathcal{C})$$

Very hard to handle...

- Probability of block error in average over all codes $\mathcal{C}$,

$$\langle p_{\mathrm{B}} \rangle \overset{\text{def}}{=} \textstyle\sum_{\mathcal{C}} p(\widehat{s} \neq s \mid \mathcal{C}) p(\mathcal{C})$$

Fortunately, this quantity is much easier to evaluate!

- The maximal block error probability for a fixed code $\mathcal{C}$,

$$p_{\mathrm{BM}}(\mathcal{C}) \overset{\text{def}}{=} \max_{s} p\left(\widehat{s} \neq s \mid s, \mathcal{C}\right)$$

It is the quantity we wish to be small!

OUR STRATEGY

1. We show that $\langle p_B \rangle$ is small
2. As $\langle p_B \rangle$ is small, it exists at least one code $\mathcal{C}$ with small p_B
3. We choose this code $\mathcal{C}$ but it could have an enormous block probability of error
4. We modify $\mathcal{C}$ $\left(\text{throwing away } 50\% \text{ of codewords, expurgation}\right)$ to guaranty that the maximal probability of block error is also small

SOMETHING WORTH REMEMBERING

In life, average properties are easy to determine in comparison with worst-case properties. . .

- For many processes, the average gives “the expected behaviour” of some process
- An average property can be useful to prove that at least one object verifies this property

Breakthrough idea of Shannon: put randomness over the code choice in order to authorize average arguments. . .

ABOUT THE DISTRIBUTION OF SYMBOLS TO ENCODE

We want:

$$p_{\mathrm{BM}}(\mathcal{C}) \overset{\text{def}}{=} \max_{s} p\left(\widehat{s} \neq s \mid s, \mathcal{C}\right) < \varepsilon$$

$\longrightarrow$ It does not depend on the s distribution!

About the distribution of symbols to encode:

To show $p_{\mathrm{BM}}(\mathcal{C}) < \varepsilon$ we will first show that $\langle p_{\mathrm{B}} \rangle$ and $p_{\mathrm{B}}(\mathcal{C})$ are $< \varepsilon$ which depends on s distribution!

Distribution of symbols:

We will suppose that s is uniformly distributed over $\left\{1, \ldots, 2^{NR'}\right\}$

$\longrightarrow$ Even if s is not chosen uniformly, it has no consequence as $p_{\mathrm{BM}}(\mathcal{C})$ does not depend on s distribution

AVERAGE PROBABILITY OF ERROR OF TYPICAL SET DECODER

The average probability of error,

$$\begin{aligned}\langle p_{\mathrm{B}} \rangle &= \sum_{\mathcal{C}} p\left(\widehat{s} \neq s \mid \mathcal{C}\right) p\left(\mathcal{C}\right) \\ &= \sum_{\mathcal{C}, s_0} p\left(\widehat{s} \neq s \mid s_0, \mathcal{C}\right) p\left(\mathcal{C}\right) p(s_0) \quad \left(s_0 \text{ and } \mathcal{C} \text{ are independent}\right) \\ &= \sum_{s_0} \frac{1}{2^{NR'}} \sum_{\mathcal{C}} p\left(\widehat{s_0} \neq s_0 \mid \mathcal{C}\right) p(\mathcal{C}) \quad \left(s \text{ is uniform}\right)\end{aligned}$$

s is encoded as $\mathbf{x}^{(s)} \in \mathcal{C}$

The distribution of $\mathcal{C} = \prod_{s=1}^{2^{NR'}} p(\mathbf{x}^{(s)})$ is invariant when permuting among coordinates $\{1, \ldots, N\}$

$\longrightarrow$ We deduce that the $\sum_{\mathcal{C}} p\left(\widehat{s_0} \neq s_0 \mid \mathcal{C}\right) p(\mathcal{C})$'s are all equal for different s_0!

Consequence:

We can write $\langle p_{\mathrm{B}} \rangle$ as $\left(\text{suppose that } s = 1 \text{ is encoded}\right)$,

$$\langle p_{\mathrm{B}} \rangle = \textstyle\sum_{\mathcal{C}} p\left(\widehat{1} \neq 1 \mid \mathcal{C}\right) p\left(\mathcal{C}\right)$$

AVERAGE PROBABILITY OF ERROR OF TYPICAL SET DECODER

$s = 1$ has been encoded as $\mathbf{x}^{(1)}$ which was transmitted and the received signal is $\mathbf{y}$

- By the jointly typical theorem,

$$p\left((\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}\right) = \delta \xrightarrow[N\to+\infty]{} 0$$

- The distribution of $\mathbf{y}$ depend only of $\mathbf{x}^{(1)}$ and the distribution of $\mathbf{x}^{(s')}$ with $s' \neq 1$ is independent of $\mathbf{y}$. By 3. of the jointly typical theorem,

$$p\left((\mathbf{x}^{(s')}, \mathbf{y}) \in J_{N\beta}\right) \leq 2^{-N(I(X,Y)-3\beta)}$$

Notice that $\{\widehat{1} \neq 1\}$ for a fixed code $\mathcal{C}$ is included in the event

$$\left\{(\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}\right\} \bigcup \left\{(\mathbf{x}^{(2)}, \mathbf{y}) \in J_{N\beta}\right\} \bigcup \cdots \bigcup \left\{(\mathbf{x}^{(2^{NR'})}, \mathbf{y}) \in J_{N\beta}\right\}$$

Then by union-bound,

$$\begin{aligned}
\langle p_{\mathrm{B}} \rangle &= \sum_{\mathcal{C}} p(\widehat{1} \neq 1 \mid \mathcal{C}) p(\mathcal{C}) \\
&\leq \sum_{\mathcal{C}} \left(p\left((\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}\right) + p\left((\mathbf{x}^{(2)}, \mathbf{y}) \in J_{N\beta}\right) + \cdots + p\left((\mathbf{x}^{(2^{NR'})}, \mathbf{y}) \in J_{N\beta}\right)\right) p(\mathcal{C}) \\
&\leq \delta + \sum_{s'=2}^{NR'} 2^{-N(I(X,Y)-3\beta)} \\
&\leq \delta + 2^{-N\left(I(X,Y)-R'-3\beta\right)}
\end{aligned}$$

CHOOSING THE RATE

$$\langle p_{\mathrm{B}} \rangle \leq \delta + 2^{-N\left(I(\mathsf{X},\mathsf{Y}) - R' - 3\beta\right)} \quad \text{where} \quad \delta \xrightarrow[N \to +\infty]{} 0$$

By choosing $R' < I(\mathsf{X},\mathsf{Y}) - 3\beta$

$$\exists N_0 \colon \forall N \geq N_0, \quad \langle p_{\mathrm{B}} \rangle < \varepsilon$$

$\longrightarrow$ We deduce:

It exists a code $\mathcal{C}$ such that $p_{\mathrm{B}}(\mathcal{C}) = p(\widehat{s} \neq s \mid \mathcal{C}) < \varepsilon$

$$\left(\text{otherwise, } \langle p_B \rangle = \textstyle\sum_{\mathcal{C}} p_{\mathrm{B}}(\mathcal{C}) p(\mathcal{C}) \geq \varepsilon \textstyle\sum_{\mathcal{C}} p(\mathcal{C}) = \varepsilon\right)$$

EXISTENCE OF A NOT TOO BAD CODE

We choose $\mathcal{C}$ such that $p_{\mathrm{B}}(\mathcal{C}) = p(\widehat{s} \neq s \mid \mathcal{C}) < \varepsilon$

We can explicitly choose this code by computing all the probabilities, but it is highly inefficient

$$\left(\text{exponential time algorithm}\ldots\right)$$

But our aim:

$$p_{\mathrm{BM}}(\mathcal{C}) = \max_{s} p\left(\widehat{s} \neq s \mid s, \mathcal{C}\right) < \varepsilon$$

$\longrightarrow$ It has no reason to be true!

EXPURGATION

We have a code $\mathcal{C}$ such that $p_{\mathrm{B}}(\mathcal{C}) = p(\widehat{s} \neq s \mid \mathcal{C}) < \varepsilon$

Fundamental Remark:

$$p(\widehat{s} \neq s \mid \mathcal{C}) = \frac{1}{2^{NR'}} \sum_{i=0}^{2^{NR'}} p(\widehat{s} \neq i \mid, s = i, \mathcal{C}) \qquad \Big(\text{s is supposed uniform}\Big)$$

$$< \varepsilon$$

We deduce,

$$\sharp \left\{ \mathrm{x}^{(i)} \in \mathcal{C} : \; p(\widehat{s_0} \neq i \mid s = i, \mathcal{C}) > 2\varepsilon \right\} \leq \tfrac{1}{2} \sharp \mathcal{C} = \tfrac{1}{2}\, 2^{NR'}$$

$$\Big(\text{otherwise } p(\widehat{s} \neq s \mid \mathcal{C}) \geq \varepsilon\Big)$$

$\longrightarrow$ We can remove half of the codewords to build a new code $\mathcal{C}^{\star}$ s.t $p_{\mathrm{BM}}(\mathcal{C}^{\star}) < 2\varepsilon$

Expurgation:

Build the new code $\mathcal{C}^{\star}$ with size $2^{NR'-1}$ by keeping half of the codewords $\mathbf{x}^{(j)} \in \mathcal{C}$ having the smallest $p(\widehat{s} \neq i \mid s = i, \mathcal{C})$. The new code has rate $R = R' - \frac{1}{N}$

CONCLUSION: CHOOSING THE CODEWORD DISTRIBUTION

$\mathcal{C}^\star$ such that $p_{\text{BM}}(\mathcal{C}^\star) < 2\varepsilon$ and with rate $R = R' - \frac{1}{N} < I(\mathsf{X}, \mathsf{Y}) - 2\varepsilon - \frac{1}{N}$

- To conclude just use $\mathsf{X}^\star$ maximizing $I(\mathsf{X}, \mathsf{Y})$, *i.e.*, $I(\mathsf{X}^\star, \mathsf{Y}) = C(Q)$

Why can we choose X as we want?

The only place were we used explicitly X was when we choose the randomness over the codes,

$$p(\mathcal{C}) = \prod_{s=1}^{2^{NR'}} p(\mathsf{x}^{(s)})$$

but it could have been $\mathsf{X}^\star$

$\longrightarrow$ We have chosen maximizing $I(\mathsf{X}, \mathsf{Y})$ to obtain the best possible rate!

It proves the existence of block codes to reliably transmit information across Q

as soon as the rate $R < C(Q)$

NOISY-CHANNEL CODING: WHAT IS IMPOSSIBLE

PROBABILITIES OF ERROR

- The probability of block error $\Big($average$\Big)$:

$$p_{\mathrm{B}} \overset{\mathrm{def}}{=} \textstyle\sum_{s_{\mathrm{in}}} p(s_{\mathrm{out}} \neq s_{\mathrm{in}} \mid s_{\mathrm{in}}) p(s_{\mathrm{in}})$$

- The maximal probability of block error $\Big($worst-case$\Big)$:

$$p_{\mathrm{BM}} \overset{\mathrm{def}}{=} \max_{s_{\mathrm{in}}} p(s_{\mathrm{out}} \neq s_{\mathrm{in}} \mid s_{\mathrm{in}})$$

We will show that: $p_{\mathrm{BM}} \to 0$ when $N \to +\infty$ implies that $R \leq C(Q)$

Given a fixed code $\mathcal{C}$:

- S distribution of s_{in} that we suppose uniform
- X^N distribution of the encoding of S into $\mathcal{C}$
- Y^N distribution of the received signals after sending X^N
- $\widehat{\mathsf{S}}$ distribution after decoding (whatever is the decoding) Y^N

$$\mathsf{S} \longrightarrow \mathsf{X}^N \longrightarrow \mathsf{Y}^N \longrightarrow \widehat{\mathsf{S}}$$

$$p(\widehat{\mathsf{S}} \neq \mathsf{S}) = p_{\mathrm{B}}$$

AN IMPORTANT REMINDER: DATA PROCESSING INEQUALITY

Notation:

$X_1 \to X_2 \to \cdots \to X_n$ denotes a Markov chain of order 1, *i.e.*,

$$p(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1) = p(X_n = x_n \mid X_{n-1} = x_{n-1})$$

In our context we have the following Markov Chain of order 1:

$$S \longrightarrow X^N \longrightarrow Y^N \longrightarrow \widehat{S}$$

Data processing inequality:

Given, $X \to Y \to Z$, then

$$I(Z, X) \leq I(Y, X)$$

FANO'S INEQUALITY

We know a random variable Y and we wish to guess the value of a correlated random variable X

$\mathsf{X} \longrightarrow \mathsf{Y} \longrightarrow \widehat{\mathsf{X}}$: where $\widehat{\mathsf{X}}$ modelizes our guess

⟶ Fano's inequality relates the probability of error in guessing the random variable X to its conditional entropy $H(\mathsf{X} \mid \mathsf{Y})$

Fano's inequality:

Let $\mathsf{X} \to \mathsf{Y} \to \widehat{\mathsf{X}}$ and $p_e \overset{\text{def}}{=} p(\widehat{\mathsf{X}} \neq \mathsf{X})$. We have,

$$H(\mathsf{X} \mid \mathsf{Y}) \leq H(\mathsf{X} \mid \widehat{\mathsf{X}}) \leq h(p_e) + p_e \log_2(\sharp\mathcal{X} - 1)$$

⟶ Lower-bound on the "probability of making a false guess" as function of conditional entropy

PROOF

Proof:

Define,

$$E = \begin{cases} 1 \text{ if } \widehat{X} \neq X \\ 0 \text{ otherwise} \end{cases}$$

By using the chain rule to expand $H(E, X \mid \widehat{X}) = H(\widehat{X}, X, E) - H(\widehat{X}) = H(\widehat{X}, E, X) - H(\widehat{X})$,

$$H(E, X \mid \widehat{X}) = H(X \mid \widehat{X}) + \underbrace{H(E \mid X, \widehat{X})}_{=0} = \underbrace{H(E \mid \widehat{X})}_{\leq h(p_e)} + \underbrace{H(X \mid E, \widehat{X})}_{\leq p_e \log_2(\sharp\mathcal{X} - 1)}$$

Since conditioning reduces the entropy, $H(E \mid \widehat{X}) \leq H(E) = h(p_e)$. The upper-bound on the last term is coming from

$$\begin{aligned} H(X \mid E, \widehat{X}) &= p(E = 0)H(X \mid \widehat{X}, E = 0) + p(E = 1)H(X \mid \widehat{X}, E = 1) \\ &\leq (1 - p_e)0 + p_e \log_2(\sharp\mathcal{X} - 1) \end{aligned}$$

where we used that knowing $\widehat{X}$ and $E = 1 \iff \widehat{X} \neq X$ and X can only take $\sharp\mathcal{X} - 1$ values. It shows the second inequality. To prove the first inequality, notice that $X \to Y \to \widehat{X}$. Therefore by the data processing inequality,

$$I(X, \widehat{X}) \leq I(X, Y) \iff H(X \mid \widehat{X}) \geq H(X \mid Y)$$

CONCLUSION

By applying Fano's inequality to our code/decoding context $\left(\mathsf{X}^N \in \mathcal{C} \text{ where } \sharp\mathcal{C} = 2^{NR}\right)$

$$\mathsf{S} \longrightarrow \mathsf{X}^N \longrightarrow \widehat{\mathsf{S}}$$

$$H(\mathsf{S} \mid \widehat{\mathsf{S}}) \leq 1 + p_{\mathrm{B}}\ NR$$

A LEMMA

Given our code/decoding context $\left(\mathsf{X}^N \in \mathcal{C} \text{ where } \sharp\mathcal{C} = 2^{NR}\right)$

$$\mathsf{S} \longrightarrow \mathsf{X}^N \longrightarrow \mathsf{Y}^N \longrightarrow \widehat{\mathsf{S}}$$

Lemma:

$$I(\mathsf{X}^N, \mathsf{Y}^N) \leq NC(Q)$$

PROOF

Proof:

By the chain rule,

$$\begin{aligned} I(X^N, Y^N) &= H(Y^N) - H(Y^N \mid X^N) \\ &= H(Y^N) - \sum_{i=1}^{N} H(Y_i \mid Y_1, \ldots, Y_{i-1}, X^N) \\ &= H(Y^N) - \sum_{i=1}^{N} H(Y_i \mid X_i) \end{aligned}$$

as by definition of the memoryless discrete channel, Y_i depends only on X_i and is conditionally independent of everything else.

$$\begin{aligned} I(X^N, Y^N) &= H(Y^N) - \sum_{i=1}^{N} H(Y_i \mid X_i) \\ &\leq \sum_{i=1}^{N} H(Y_i) - \sum_{i=1}^{N} H(Y_i \mid X_i) \\ &= \sum_{i=1}^{N} I(X_i, Y_i) \\ &\leq NC(Q) \end{aligned}$$

CONCLUDING THE PROOF OF SHANNON'S THEOREM

$$\left(p_B = p(\widehat{\mathsf{S}} \neq \mathsf{S})\right)$$

Proof:

We have $\mathsf{S} \longrightarrow \mathsf{X}^N \longrightarrow \mathsf{Y}^N \longrightarrow \widehat{\mathsf{S}}$ where S is uniform over $\{1, \dots, 2^{NR}\}$. Therefore,

$$\begin{aligned} NR &= H(\mathsf{S}) \\ &= H(\mathsf{S} \mid \widehat{\mathsf{S}}) + I(\mathsf{S}, \widehat{\mathsf{S}}) \\ &\leq 1 + p_{\mathrm{B}}\ NR + I(\mathsf{S}, \widehat{\mathsf{S}}) \quad \left(\text{Fano's inequality}\right) \\ &\leq 1 + p_{\mathrm{B}}\ NR + I(\mathsf{X}^N, \mathsf{Y}^N) \quad \left(\text{data processing inequality}\right) \\ &\leq 1 + p_{\mathrm{B}}\ NR + NC(Q) \end{aligned}$$

showing that

$$p_{\mathrm{B}} \geq 1 - \frac{C(Q)}{R} - \frac{1}{NR}$$

But $p_{\mathrm{BM}} \to 0$ by assumption. In particular, $p_{\mathrm{B}} \to 0$. We deduce from the previous inequality that $R \leq C(Q)$. It concludes the proof of Shannon noisy-channel coding theorem

Remark:

We supposed that S is uniform $\left(\text{over the messages}\right)$ without loss of generality. It can be done as above we only need to use that $H(\mathsf{S}) \leq NR$

CONCLUSION

COMBINING COMPRESSION AND NOISY-CHANNEL CODING

It is coming the time to combine everything we have done. . .

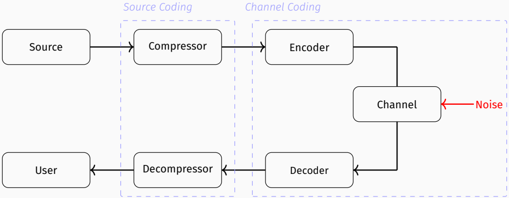
- To compress: $R > H$
- To send over a noisy channel: $R < C$

But are we loosing something by the two stages process: first compress and then encode to send across a noisy channel?

⟶ We can prove that the two stage method is as good as any other method of transmitting information!

Consequence: consider the deign of an encoding system as source coding and channel coding

1. Design source code for the most efficient representation
2. Separately and independently, design channel code appropriate for the channel



NOISY-CHANNEL CODING THEOREM AND CODING IN PRACTICE

Communication over a noisy channel is possible! But our proof gave an highly non-efficient algorithm. . .

$\longrightarrow$ In Lecture 7 we start to show how to design efficient encoding-decoding algorithms!

$\big($in particular, linear codes$\big)$

Is our channel model relevant?

Berlekamp (1980): design encoding-decoding algorithms and plot their performance on a variety of idealized channels as a function of the noise. These charts can be shown to the customer, who can choose among the systems on offer without having to specify what he really thinks his channel is like!

AN IMPORTANT CLASS OF CHANNELS: SYMMETRIC CHANNELS

Symmetric Discrete Memoryless Channel (SDMC):

A discrete memoryless channel is said symmetric if the set of outputs can be partitioned into subsets such that for each subset the matrix of transition probabilities has the property that each row is a permutation of each other row and each column is a permutation of each other column.

An example:

The channel,

$$\begin{array}{cc} p(y=0 \mid x=0)=0.7 & p(y=0 \mid x=1)=0.1 \\ p(y=\perp \mid x=0)=0.2 & p(y=\perp \mid x=1)=0.2 \\ p(y=1 \mid x=0)=0.1 & p(y=1 \mid x=1)=0.7 \end{array}$$

is symmetric. Partition the outputs according to $\{0,1\}$ and $\{\perp\}$.

$$\begin{array}{cc} \hline p(y=0 \mid x=0)=0.7 & p(y=0 \mid x=1)=0.1 \\ p(y=1 \mid x=0)=0.1 & p(y=1 \mid x=1)=0.7 \\ \hline p(y=\perp \mid x=0)=0.2 & p(y=\perp \mid x=1)=0.2 \\ \hline \end{array}$$

SYMMETRIC CHANNELS AND LINEAR CODES

In Lecture 7 we will introduce a sub-class of block-codes: linear codes

⟶ Linear codes admit compact representations and encoding algorithm

But do linear codes reach the capacity of any memoryless discrete channel?

⟶ Linear codes reach the capacity of any symmetric discrete memoryless channel!

LAST LAST CONCLUSION

Computing the capacity of a given channel is usually a hard problem...

$\longrightarrow$ But an important case for which we know the capacity:

Weakly Symmetric Channel:

A SDMC is said to be weakly symmetric if every row of the transition matrix $p(\cdot \mid x)$ is a permutation of every other rows and all the columns sums $\sum_y p(y \mid x)$ are equal

Capacity of weakly symmetric channel:

For a weakly symmetric channel Q,

$$C(Q) = \log_2 \sharp\mathcal{Y} - H(\text{row of transition matrix})$$

and this is achieved by a uniform distribution on the input alphabet

EXERCISE SESSION