# Information Theory

# Chapter IV: Discrete transmission channels

# What are they?

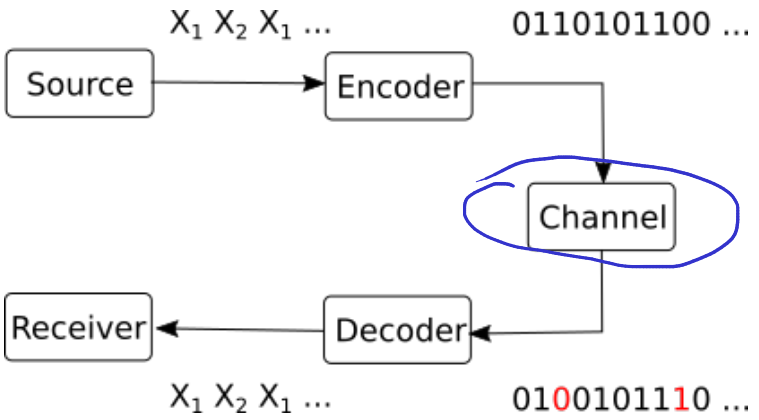
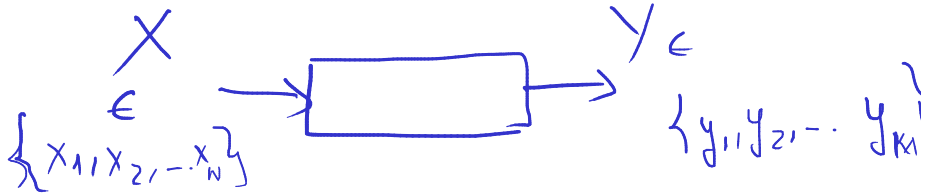



Figure 1: Communication system

- A device that transmits data from one place to another
- The data undergoes **distortions** / **errors**
- We consider that transmission is instantaneous

# How do they work?

- A random variable $X \in \{x_1, x_2, ...\}$ is put at the input of the channel
- A random variable $Y \in \{y_1, y_2, ...\}$ appears immediately at the output of the channel
  - $Y$ is related to $X$




- The receiver wants to find $X$, but can see only $Y$

{0, 1}

{0, 1}

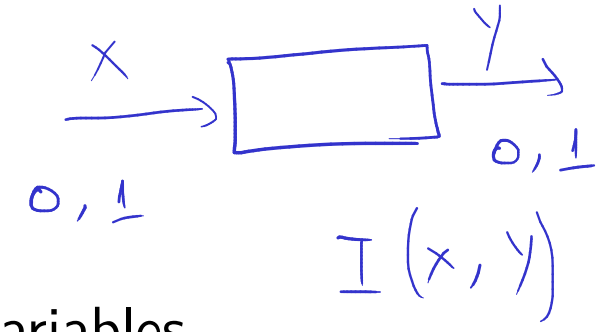
Naming:

- The inputs $\{x_1, x_2, ...\}$ and outputs $\{y_1, y_2, ...\}$ are called **symbols**
- Symbols $\neq$ messages $s_i$ from the source S
  - The encoder might convert $s_i$ to a different representation
  - Example: source messages = characters, but channel symbols = 0/1 (encoder converts characters to binary)

# What do we want?

- A successful communication = deduce the $X$ which was sent from the $Y$ that was received
- We are interested in **deducing X when knowing just Y**
- Main topic: How much does knowing $Y$ tell us about $X$?
  - Depends on the relation between them
  - Is the same as how much $X$ tells us about $Y$ (symmetrical)

# Probabilistic description




From a probabilistic point of view:

- A system of two related random variables
  - Input random variable $X \in \{x_1, x_2, ...\}$
  - Output random variable $Y \in \{y_1, y_2, ...\}$
  - It doesn't matter that one is *input* and other is *output*, we just care about the relation between the two random variables

X, Y not-independent

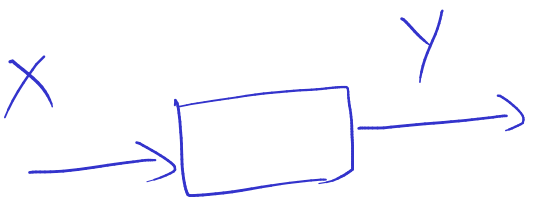
- $X$ and $Y$ are *related* probabilistically, but still *random* (because of noise / errors / distortions)
  - All the probabilities are known

- We need to analyze the relation of $X$ with $Y$

# Intuitive examples

- Binary channel with errors
  - Send 0's and 1's, receive 0's and 1's, but with errors


- Pipe
  - Send colored balls over the pipe, but someone may be re-painting them
- Grandma calling!
  - She says *"cat"* / *"hat"* / *"pet"*, but sometimes you hear her wrong
- Living near stadium
  - You don't actually see the game, but try to deduce the score from the shouts you hear



# Nomenclature

We only deal with discrete memoryless stationary channels

- Discrete: number of input and output symbols is finite
- Memoryless: the output symbol depends only on the current input symbol
- Stationary: the probabilities involved do not change in time

# Systems of two random variables

- Two random variables: $X = \{x_1, x_2, ...\}$, $Y = \{y_1, y_2, ...\}$.
- Example: throw a dice (X) and a coin (Y) simultaneously
- How to describe this system?

A single joint information source:

$\cap$ = "and"


$P(x_i \cap y_j)$

$$X \cap Y : \begin{pmatrix} x_1 \cap y_1 & x_1 \cap y_2 & \dots & x_i \cap y_j \\ p(x_1 \cap y_1) & p(x_1 \cap y_2) & \dots & p(x_i \cap y_j) \end{pmatrix}$$

Arrange in a nicer form (table):

| | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | ... | ... | ... |
| $x_2$ | ... | ... | ... |
| $x_3$ | ... | ... | ... |

| X \ Y | 0 | 1 |
|---|---|---|
| 0 | 0.3 | 0.01 |
| 1 | 0.05 | 0.64 |

$P(X=1 \cap Y=0)$

- Elements of the table: $p(x_i \cap y_j)$

# Joint probability matrix

The table constitutes the **joint probability matrix**:

$$P(X,Y) = \begin{bmatrix} p(x_1 \cap y_1) & p(x_1 \cap y_2) & \cdots & p(x_1 \cap y_M) \\ p(x_2 \cap y_1) & p(x_2 \cap y_2) & \cdots & p(x_2 \cap y_M) \\ \vdots & \vdots & \cdots & \vdots \\ p(x_N \cap y_1) & p(x_N \cap y_2) & \cdots & p(x_N \cap y_M) \end{bmatrix}$$

$$\sum_i \sum_j p(x_i \cap y_j) = 1$$

- This matrix completely defines the two-variable system
- This matrix completely defines the communication process

# Joint entropy

$$P(x,y) = \begin{bmatrix} 0.3 & 0.01 \\ 0.05 & 0.64 \end{bmatrix}$$

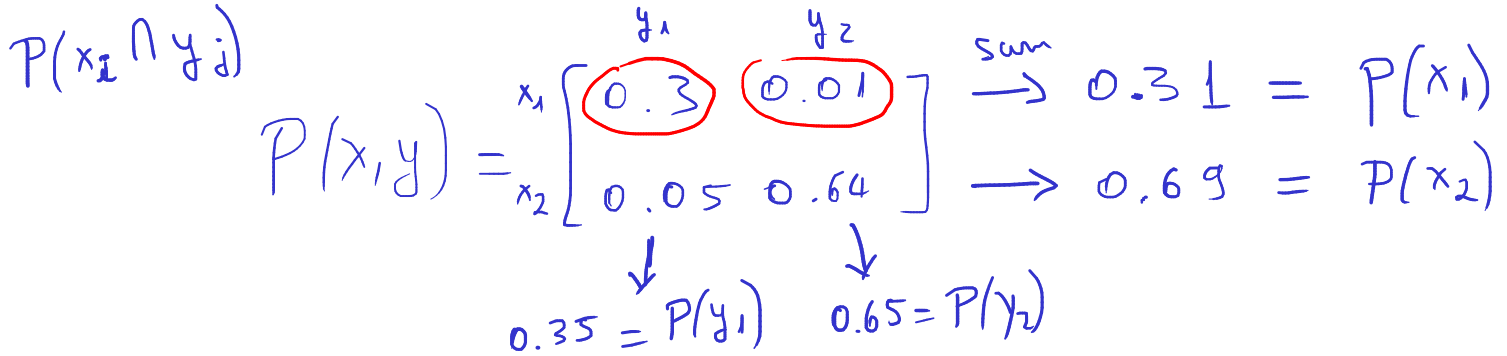
$$H(x,y) = -0.3 \log 0.3 - 0.01 \log 0.01 - 0.05 \log 0.05 - 0.64 \log 0.64$$

- The distribution $X \cap Y$ determines the **joint entropy**:

$$H(X,Y) = -\sum_i \sum_j p(x_i \cap y_j) \cdot \log(p(x_i \cap y_j))$$

- This is the global entropy of the system (knowing the input and the output)

# Marginal distributions




- $p(x_i) = \sum_j p(x_i \cap y_j)$ = sum of row $i$ from P(X,Y)
- $p(y_j) = \sum_i p(x_i \cap y_j)$ = sum of column $j$ from P(X,Y)
- The distributions $p(x)$ and $p(y)$ are called **marginal distributions** ("summed along the margins")

# Examples [marginal distributions not enough]

Marginal distributions don't tell everything about the system:

- Example 1:

$$H(X) = -0.3 \log 0.3 - 0.7 \log 0.7$$
$$H(Y) = -0.3 \log 0.3 - 0.7 \log 0.7$$

$$P(X,Y) = \begin{matrix} x_1 \\ x_2 \end{matrix} \begin{matrix} y_1 & y_2 \end{matrix} \begin{bmatrix} 0.3 & 0 \\ 0 & 0.7 \end{bmatrix} \rightarrow p(x_1) = 0.3, \; p(x_2) = 0.7 \Big\} H(X)$$

$$\downarrow \qquad \downarrow$$
$$p(y_1) = 0.3 \qquad 0.7 = p(y_2) \quad \Big\} H(Y)$$

- Example 2:

$$P(X,Y) = \begin{matrix} x_1 \\ x_2 \end{matrix} \begin{matrix} y_1 & y_2 \end{matrix} \begin{bmatrix} 0.15 & 0.15 \\ 0.15 & 0.55 \end{bmatrix} \rightarrow p(x_1) = 0.3, \; p(x_2) = 0.7$$

$$\downarrow 0.3 \qquad \downarrow 0.7$$

- Both have identical $p(x)$ and $p(y)$, but are completely different
- Which one is better for a transmission?
- Marginal distribution are useful, but not enough. Essential is the *relation* between X and Y.

# Bayes formula

$$p(A \cap B) = p(A) \cdot p(B|A) = P(B) \cdot p(A|B)$$

$P(B|A) = p(B \text{ if } A)$

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

- "The conditional probability of B **given A**" (i.e. given that event A happened)
- Examples: listen to the lecture

When A and B are independent events:

$$p(A \cap B) = p(A)p(B)$$

$$p(B|A) = p(B)$$

- The fact that event $A$ happened doesn't influence $B$ at all

# Three examples

$$P(\text{shoot} \cap \text{goal}) = P(\text{shoot}) \cdot P(\text{goal} \mid \text{shoot})$$

"AND" "IF"

$$P(\text{sad} \cap \text{poor}) = P(\text{poor}) \cdot P(\text{sad} \mid \text{poor})$$

Examples to help you remember conditional probabilities

- Gambler's paradox
- CNN: Crippled cruise ship returns; passengers happy to be back

23 Red $\quad P(24 \text{ Reds}) = \text{very small}$

$P(23 \text{ Red} \ \& \ 1 \text{ Black}) = - // -$

$$P(23 \text{ Red} \ \& \ 1 \text{ Black} \mid 23 \text{ Red}) = \frac{1}{2} = \frac{P(23 \text{ Red} \cap 1 \text{ Black})}{P(23 \text{ Reds})}$$

$P(\text{Kid2} \mid \text{Kid1})$

# Channel matrix

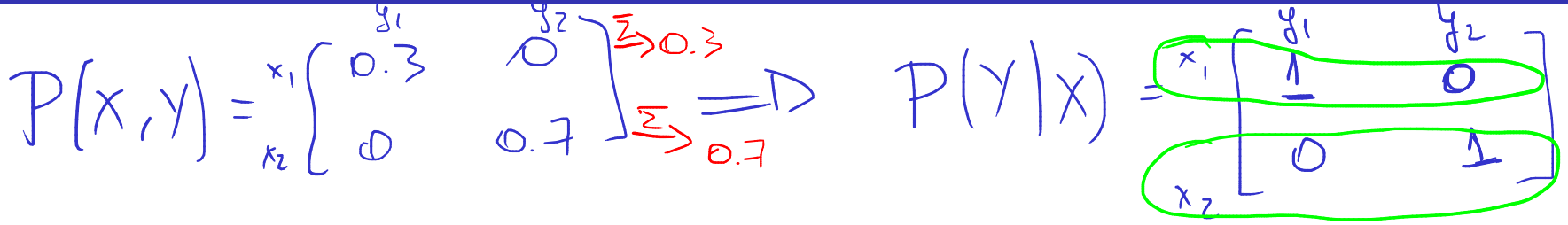
$$P(y_1|x_1) = \frac{P(x_1 \cap y_1)}{P(x_1)}$$

Noise (or channel) matrix:

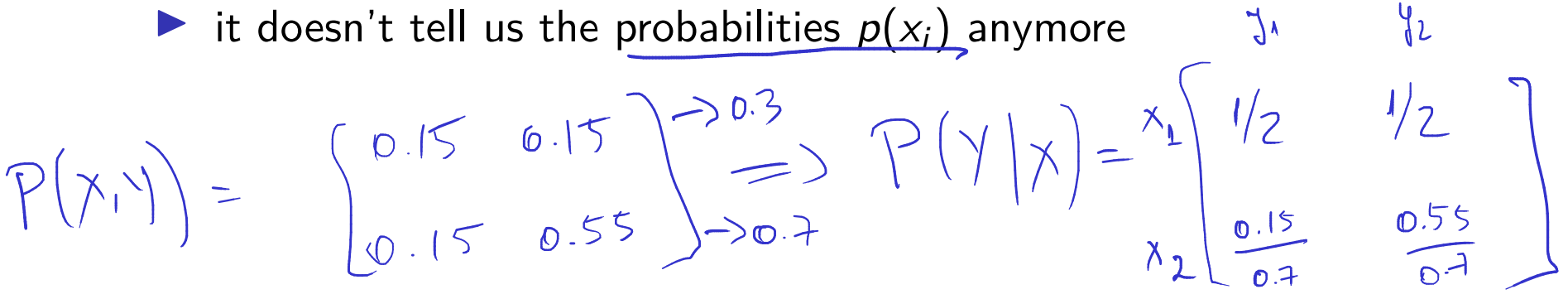
$$P(Y|X) = \begin{array}{c} \\ x_1 \\ x_2 \\ \\ x_N \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & & y_M \end{array} \\ \begin{bmatrix} p(y_1|x_1) & p(y_2|x_1) & \cdots & p(y_M|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \cdots & p(y_M|x_2) \\ \vdots & \vdots & \cdots & \vdots \\ p(y_1|x_N) & p(y_2|x_N) & \cdots & p(y_M|x_N) \end{bmatrix} \end{array} \begin{array}{l} \\ \rightarrow H(Y|x_1) \\ \rightarrow H(Y|x_2) \\ \\ \vdots \end{array}$$

- Defines the probability of an output **given an input**
- Each row = a separate distribution that indicates the probability of the outputs **if the input is** $x_i$
- The sum of each row is 1 (there must be some output if the input is $x_i$

# Relation of channel matrix and joint probability matrix



- $P(Y|X)$ is obtained from $P(X,Y)$ by dividing every row to its sum $(p(x_i))$
- This is known as *normalization* of rows
- $P(X,Y)$ can be obtained back from $P(Y|X)$ by multiplying each row with $p(x_i)$
- $P(Y|X)$ contains less information than $P(X,Y)$
  - it doesn't tell us the probabilities $p(x_i)$ anymore



# Definition of a discrete transmission channel

$P(0) = 40\%$

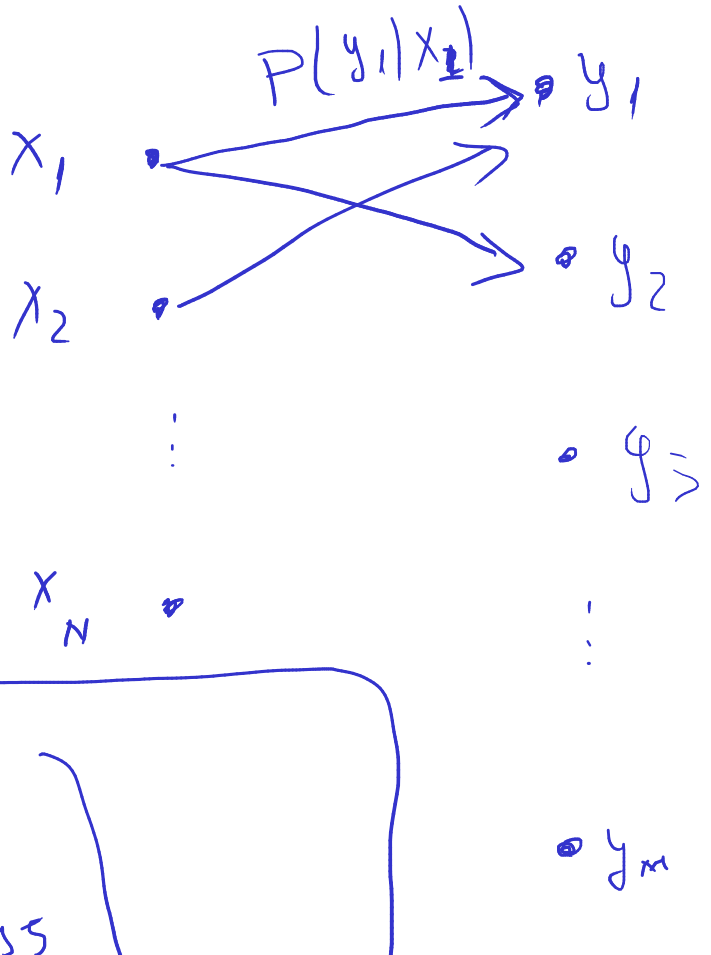
$P(1) = 60\%$

**Definition**: A discrete transmission channel is defined by three items:

1. The input alphabet $X = \{x_1, x_2, \ldots\}$
2. The output alphabet $Y = \{y_1, y_2, \ldots\}$
3. The noise (channel) matrix $P(Y|X)$ which defines the conditional probabilities of the outputs $y_j$ for every possible input $x_i$

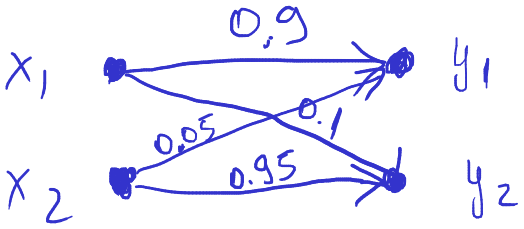
$$P(Y|X) = \begin{array}{c} \\ x_1=0 \\ x_2=1 \end{array} \begin{array}{c} \begin{array}{cc} y_1=0 & y_2=1 \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix} \end{array}$$

# Graphical representation of a channel

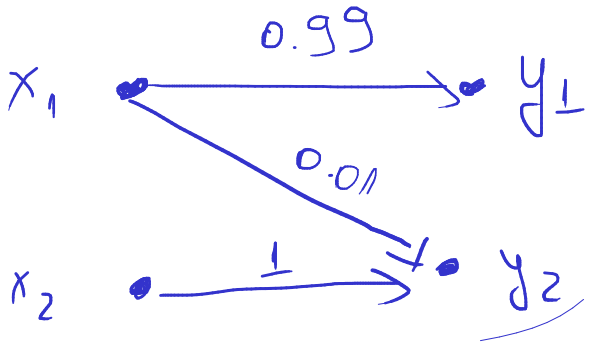



- Nice picture with arrows :)

$$P(Y|X) = \begin{matrix} \\ x_1=0 \\ x_2=1 \end{matrix} \begin{matrix} y_1=0 & y_2=1 \end{matrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix}$$



# Intuitive examples




- Postal service
- Play and win the lottery
  - $+$ funny joke

# Conditional entropy H(Y|X) (mean error)

- Since each row in $P(Y|X)$ is a distribution, each row has an entropy
- Entropy of row $x_i$:

$$H(Y|x_i) = -\sum_j p(y_j|x_i) \log(p(y_j|x_i))$$

- $H(Y|x_i)$ = *"The uncertainty of the output symbol when the input symbol is $x_i$"*
- Example: lottery

$$P(Y|X) = \begin{array}{c|cc} & y_1 & y_2 \\ x_1 & 0.9 & 0.1 \\ x_2 & 0.05 & 0.95 \end{array}$$

$$H(Y|x_1) = -0.9 \log 0.9 - 0.1 \log 0.1$$

$$H(Y|x_2) = -0.05 \log 0.05 - 0.95 \log 0.95$$

# Conditional entropy H(Y|X) (mean error)

$H(Y|X) = p(x_1) \cdot H(Y|x_1) + p(x_2) \cdot H(Y|x_2) + \ldots$

- There may be a different value $H(Y|x_i)$ for every $x_i$
- Compute the average over all $x_i$:

$$H(Y|X) = \sum_i p(x_i) H(Y|x_i) \quad = \text{average of } H(Y|x_i)$$

$$= -\sum_i \sum_j p(x_i) p(y_j|x_i) \log(p(y_j|x_i))$$

$$= -\sum_i \sum_j p(x_i \cap y_j) \log(p(y_j|x_i))$$

- $H(Y|X)$ = **"The uncertainty of the output symbol when we know the input symbol"** (any input, in general)
- Also known as **average error**

# Equivocation matrix

$$P(x_i \mid y_j) = \frac{P(x_i \cap y_j)}{P(y_j)}$$

Equivocation matrix:

$$P(X|Y) = \begin{bmatrix} p(x_1|y_1) & p(x_1|y_2) & \cdots & p(x_1|y_M) \\ p(x_2|y_1) & p(x_2|y_2) & \cdots & p(x_2|y_M) \\ \vdots & \vdots & \cdots & \vdots \\ p(x_N|y_1) & p(x_N|y_2) & \cdots & p(x_N|y_M) \end{bmatrix}$$

- Defines the probability of an input **given an output**
- Each column = a separate distribution that indicates the probability of the inputs **if the output is** $y_j$
- The sum of each column is 1 (there must be some input if the output is $y_j$

# Relation of equivocation matrix and joint probability matrix

$P(x_i \mid y_j)$

- $P(X|Y)$ is obtained from $P(X, Y)$ by dividing every column to its sum ($p(y_j)$)
- This is known as *normalization* of columns
- $P(X, Y)$ can be obtained back from $P(X|Y)$ by multiplying each column with $p(y_j)$
- $P(X|Y)$ contains less information than $P(X, Y)$
  - it doesn't tell us the probabilities $p(y_i)$ anymore

$$P(X,Y) = \begin{matrix} & y_1 & y_2 \\ x_1 & 0.15 & 0.15 \\ x_2 & 0.15 & 0.55 \end{matrix}$$

$P(y_1) = 0.3 \quad P(y_2) = 0.7$

$$\Rightarrow P(X|Y) = \begin{matrix} & y_1 & y_2 \\ x_1 & 1/2 & 0.15/0.7 \\ x_2 & 1/2 & 0.55/0.7 \end{matrix}$$

$H(X \mid y_1) = 1 \qquad H(X \mid y_2) =$

# Conditional entropy H(X|Y) (equivocation)

$H(X|y_j)$

- Since each column is a distribution, each column has an entropy
- Entropy of column $y_j$:

entropy of column $y_j$ from $P(x/y)$

$$H(X|y_j) = -\sum_i p(x_i|y_j) \log(p(x_i|y_j))$$

- $H(X|y_j) =$ *"The uncertainty of the input symbol when the output symbol is $y_j$"*

# Conditional entropy H(X|Y) (equivocation)

$H(X|Y) = p(y_1) \cdot H(X|y_1) + p(y_2) \cdot H(X|y_2) + \ldots$

- A different $H(X|y_j)$ for every $y_j$
- Compute the average over all $y_j$:

$$H(X|Y) = \sum_j p(y_j) H(X|y_j) \quad = \text{average of } H(X|y_j)$$

$$= -\sum_i \sum_j p(y_j) p(x_i|y_j) \log(p(x_i|y_j))$$

$$= -\sum_i \sum_j p(x_i \cap y_j) \log(p(x_i|y_j))$$

- **"The uncertainty of the input symbol when we know the output symbol"** (any output, in general)
- Also known as **equivocation**
- Should be small for a good communication

# The big picture

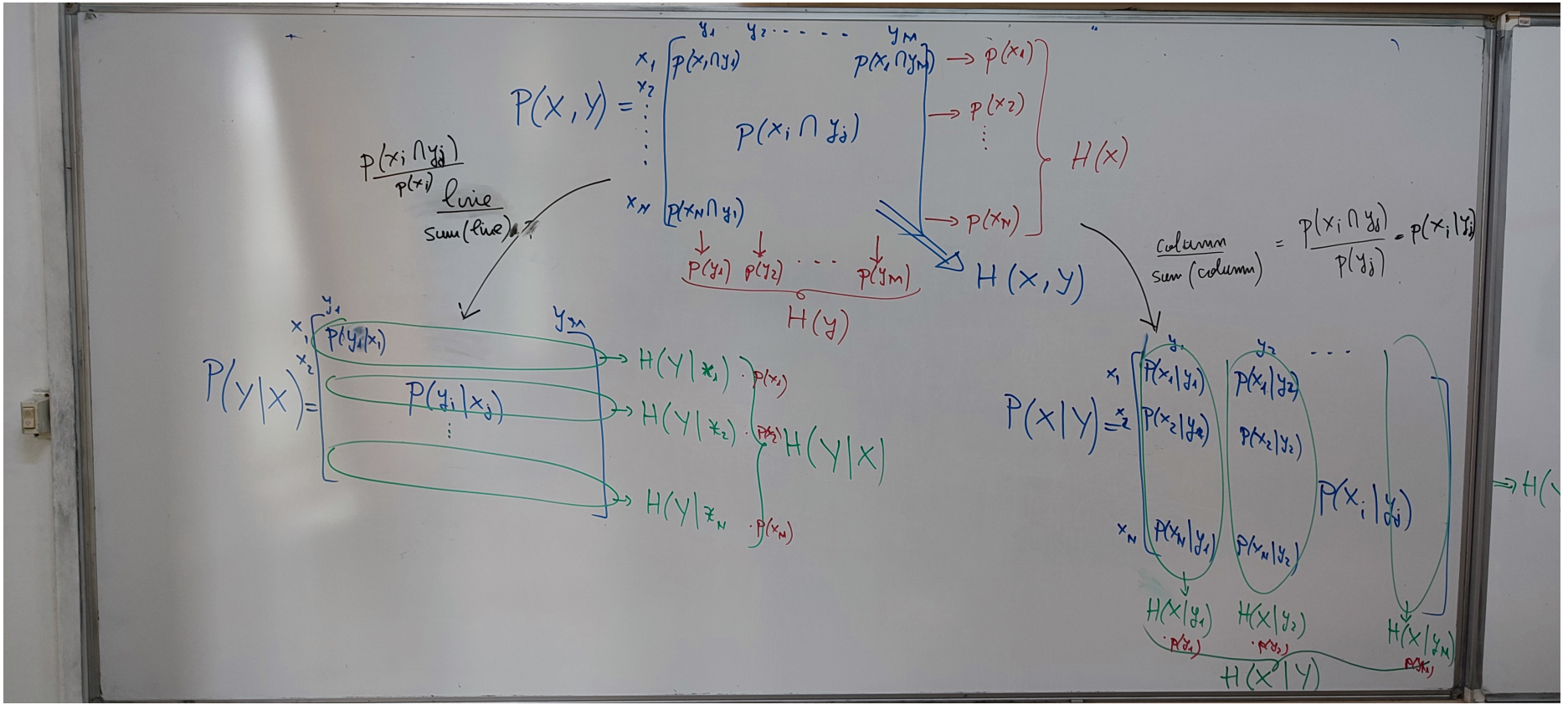



Figure 2: The big picture

# Properties of conditional entropies

For a general system with two random variables $X$ and $Y$:

- Conditioning always reduces entropy:

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

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

(knowing something cannot harm)

- If the variables are independent:

$$H(X|Y) = H(X)$$

$$H(Y|X) = H(Y)$$

(knowing the second variable does not help at all)

# Relations between the informational measures

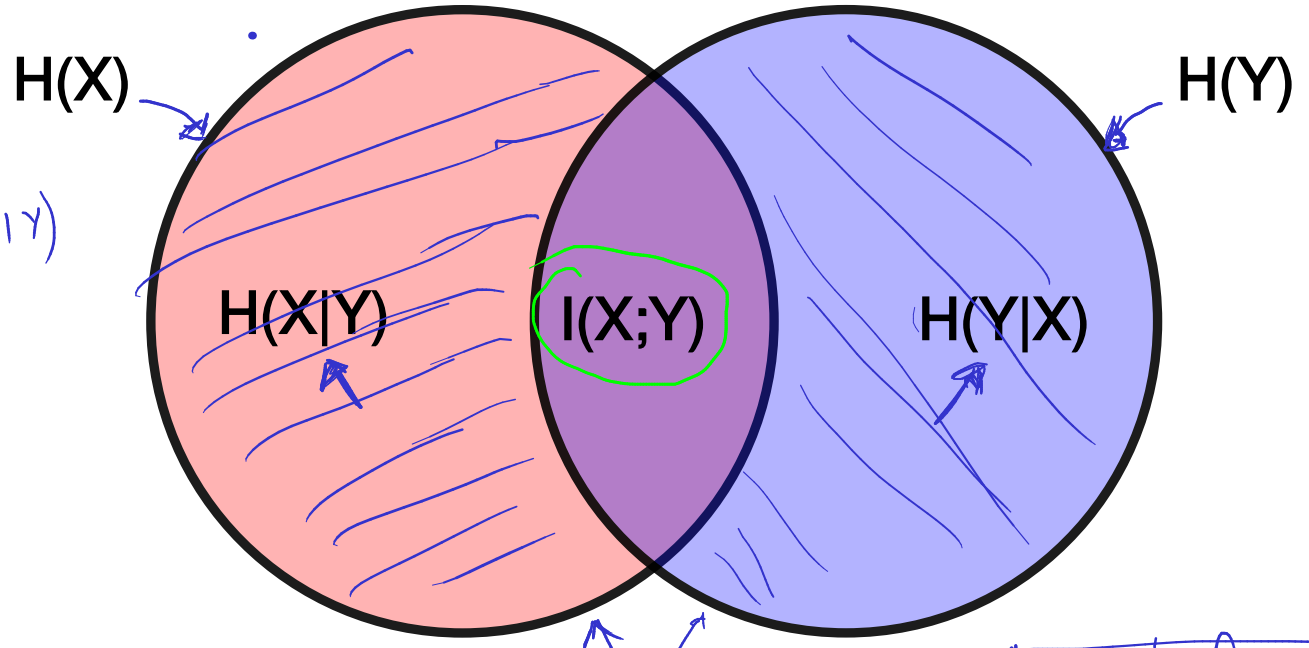



Figure 3: Relations between the informational measures

- *image from Wikipedia*

# Relations between the informational measures

- Six quantities: $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$, $I(X,Y)$
- All relations on the picture are valid relations:

$$H(X,Y) = H(X) + H(Y) - I(X,Y)$$

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

...

- If know three, can find the other three
- Simplest to find first H(X), H(Y), H(X,Y) —> then find others

# Mutual information I(X,Y)

- Mutual information I(X,Y) = the average information that one variable has about the other
- Mutual information I(X,Y) = the average information that is transmitted on the channel
- Consider a communication channel with $X$ as input and $Y$ as output:
  - We are the receiver and we want to find out the $X$
  - When we don't know the output: H(X)
  - When we know the output: H(X|Y)
- How much information was transmitted?
  - Reduction of uncertainty:

$$I(X, Y) = H(X) - H(X|Y)$$

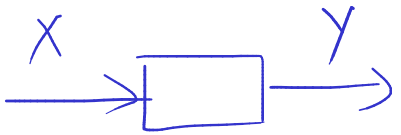
Uncert. of X without any communication

Uncert. of X if you know the output Y

# Mutual information I(X,Y)

$$
\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= -\sum_i p(x_i)\log(p(x_i)) + \sum_i \sum_j p(x_i \cap y_j)\log(p(x_i|y_j)) \\
&= -\sum_i \sum_j p(x_i \cap y_j)\log(p(x_i)) + \sum_i \sum_j p(x_i \cap y_j)\log(p(x_i|y_j)) \\
&= \sum_i \sum_j p(x_i \cap y_j)\log\left(\frac{p(x_i|y_j)}{p(x_i)}\right) \cdot p(y_j) \cdot p(y_j) \\
I(X,Y) &= \sum_i \sum_j p(x_i \cap y_j)\log\left(\frac{p(x_i \cap y_j)}{p(x_i)p(y_j)}\right)
\end{aligned}
$$

# Properties of mutual information




Mutual information $I(X, Y)$ is:

- commutative: $I(X, Y) = I(Y, X)$
- non-negative: $I(X, Y) \geq 0$
- a special case of the Kullback–Leibler distance (relative entropy distance)

$D_{KL} =$

# Relation to Kullback-Leibler distance

- $I(X, Y)$ is a special case of the Kullback-Leibler distance

$$D_{KL}(P||Q) = \sum_i p(s_i) \log\left(\frac{p(s_i)}{q(s_i)}\right)$$

- In our case, the distributions are:
  - $p(s_i) = p(x_i \cap y_j)$ = joint distribution of $X$ and $Y$ our system
  - $q(s_i) = p(x_i) \cdot p(y_j)$ = joint distribution when $X$ and $Y$ are independent

$$I(X, Y) = D_{KL}(p(x_i \cap y_j)||p(x_i) \cdot p(y_j))$$

- Interpretation
  - When $X$ and $Y$ are independent, mutual information $I(X, Y) = 0$
  - Our mutual information = how far away are from being independent
  - Example: height of a point = how far is it from the point of 0 height

# Types of communication channels

$$P(X,Y) = \begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} 0 & 0.2 & 0 \\ 0.3 & 0 & 0.5 \end{bmatrix} \end{array}$$

joint

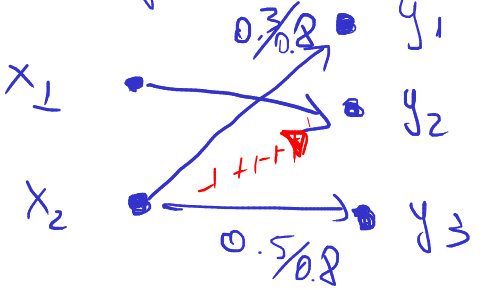
$$P(Y|X) = \begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ \frac{0.3}{0.8} & 0 & \frac{0.5}{0.8} \end{bmatrix} \end{array}$$

channel

1. Channels with zero equivocation

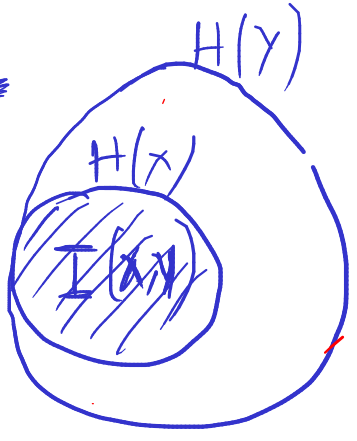
$$H(X|Y) = 0$$

Equivocation

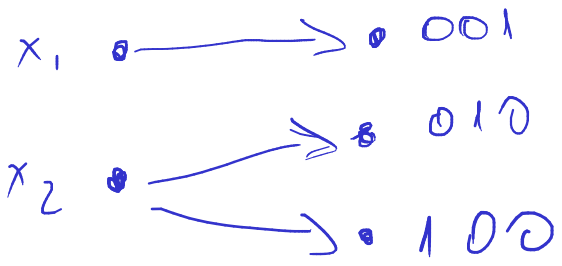



- Each column of the noise (channel) matrix contains only one non-zero value
- No doubts on the input symbols when the output symbols are known
- All input information is transmitted

$$I(X,Y) = H(X)$$



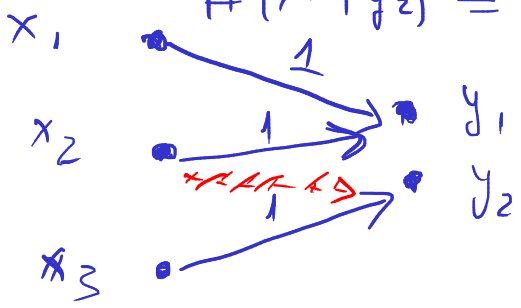

- Example: codewords...



# Types of communication channels

$$P(Y|X) = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

$$H(X|Y) \neq 0$$
$$H(X|y_1) \neq 0$$
$$H(X|y_2) = 0$$

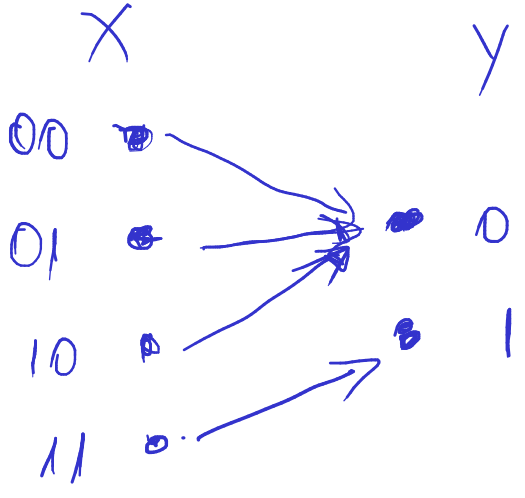



2. Channels with zero mean error

$$H(Y|X) = 0$$

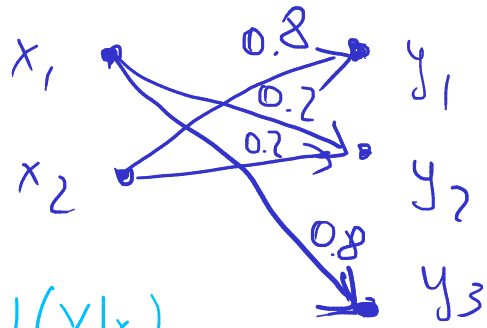
- Each row of the noise (channel) matrix contains only one non-zero value
- No doubts on the output symbols when the input symbols are known
- *The converse is not necessary true!*

- Example: AND gate



# Types of communication channels

$$P(Y|X) = \begin{array}{c|ccc} & y_1 & y_2 & y_3 \\ x_1 & 0 & 0.2 & 0.8 \\ x_2 & 0.8 & 0.2 & 0 \end{array}$$




3. Channels uniform with respect to the input

$$H(Y|x_i) = same$$

$H(Y|x_1)$
$H(Y|x_2)$
= the same value

$$H(Y|X) = H(Y|x_1) \cdot p(x_1) + H(Y|x_2) \cdot p(x_2)$$

channel

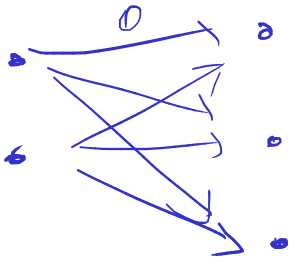
- Each row of noise matrix contains the same values, possibly in different order
- $H(Y|x_i) = same = H(Y|X)$
- $H(Y|X)$ does not depend on the actual probabilities $p(x_i)$

anything

# Types of communication channels

this is also symmetric

$$P(Y|X) = \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} \begin{matrix} y_1 & y_2 & y_3 \end{matrix} \\ \begin{bmatrix} 0.2 & 0.2 & 0.6 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \end{bmatrix} \end{matrix} \begin{matrix} \\ \to 1 \\ \to 1 \\ \to 1 \end{matrix}$$



4. Channels uniform with respect to the output

$H(X|y_1)$ $H(X|y_2)$ $H(X|y_j)$ $\Rightarrow$ $H(X|Y)$ does not dep. on $P(y_j)$

- Each column of noise matrix contains the same values, possibly in different order

$P(y_1) = \dots P(y_N)$

- If the input symbols are equiprobable, the output symbols are also equiprobable

$P(x_1) = \dots = P(x_N)$

$P(X|Y)$

- Attention:

$$H(X|y_j) \neq same!$$

A better example:

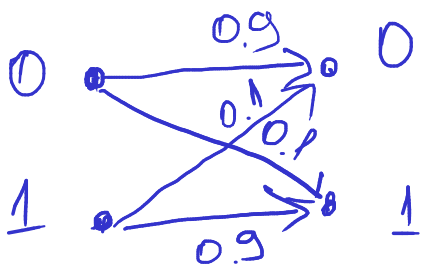
$$P(Y|X) = \left\{ \qquad \right\}$$

# Types of communication channels

$$P(Y|X) = \begin{bmatrix} 0.2 & 0.2 & 0.6 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \end{bmatrix}$$

5. Symmetric channels
   - Uniform with respect to the input and to the output
   - Example: binary symmetric channel

BSC: $P(Y|X) = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \end{matrix}$

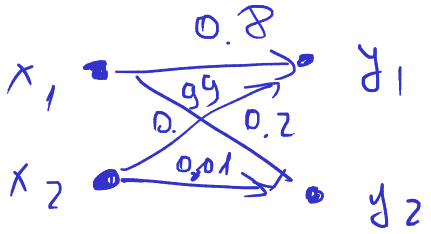



error probability

# Input probabilities are important

$p(x_1) = 0.3$
$p(x_2) = 0.7$

$p(x_1) = 0.5$
$p(x_2) = 0.5$



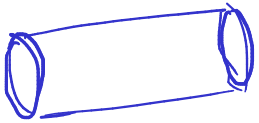

C

- Suppose we have a channel defined by P(Y|X)
- I(X,Y) **depends on the input probabilities** $p(x_i)$
  - For some distribution $p(x_i)$, we get a value of I(X,Y)
  - For a different distribution $p(x_i)$, we get a different I(X,Y)
- We want I(X,Y) to be as large as possible
- Questions:
  - what is the largest possible value of I(X,Y) (depending on $p(x_i)$)?
  - For what distribution $p(x_i)$?

# Channel capacity

- What is the maximum information I(X,Y) we can transmit on a certain channel?
- **Definition:** the **information capacity of a channel** is the maximum value of the mutual information, where the maximization is done over the input probabilities $p(x_i)$

$$C = \max_{p(x_i)} I(X, Y)$$



- i.e. the maximum mutual information we can obtain if we are allowed to choose $p(x_i)$ as we want
- Use together with definition of $I(X, Y)$:

$$C = \max_{p(x_i)} \left(H(Y) - H(Y|X)\right)$$

$$C = \max_{p(x_i)} \left(H(X) - H(X|Y)\right)$$

# What channel capacity means

- Channel capacity is the maximum information we can transmit on a channel, on average, with one symbol
- One of the most important notions in information theory
- Its importance comes from Shannon's second theorem (noisy channel theorem)
- It allows us to compare channels

STOP HERE

2022 - 2023

# Preview of the channel coding theorem

- For transmission with no errors, we use **error coding** of data before transmission
- How error coding usually works:
  - For each $k$ symbols of data, coder appends additional $m$ symbols, computed via some coding algorithm
  - All of them are sent on the channel
  - The decoder detects/corrects errors based on the additional $m$ bits
- Coding rate:

$$R = \frac{k}{k+m}$$

  - stronger protection = bigger $m$ = less efficient
  - weaker protection = smaller $m$ = more efficient

# Preview of the channel coding theorem

- A rate is called **achievable** for a channel if, for that rate, there exists a coding and decoding algorithm guaranteed to correct all possible errors on the channel

**Shannon's noisy channel coding theorem (second theorem)**

For a given channel, all rates below capacity $R < C$ are achievable. All rates above capacity, $R > C$, are not achievable.

# Channel coding theorem explained

In layman terms:

- For all coding rates $R < C$, there is a way to recover the transmitted data perfectly (decoding algorithm will detect and correct all errors)
- For all coding rates $R > C$, there is no way to recover the transmitted data perfectly

Example:

- Send binary digits (0,1) on a channel with capacity 0.7 bits/message
- There exists coding schemes with R $<$ 0.7 that allow perfect recovery
  - i.e. for every 7 bits of data coding adds 3 or more bits, on average $=>$ $R = \frac{7}{7+3}$
- With less than 3 bits for every 7 bits of data $=>$ impossible to recover all the data

# Efficiency and redundancy

- Efficiency of a channel:

$$\eta_C = \frac{I(X,Y)}{C}$$

- Absolute redundancy of a channel:

$$R_C = C - I(X,Y)$$

- Relative redundancy of a channel:

$$\rho_C = \frac{R_C}{C} = 1 - \frac{I(X,Y)}{C} = 1 - \eta_C$$

# Computing the capacity

- Tricks for easier computation of the capacity
- Channel is uniform with respect to the input:
  - $H(Y|X)$ does not depend on the actual probabilities $p(x_i)$
  - $C = \max_{p(x_i)} I(X,Y) = \max_{p(x_i)} (H(Y) - H(Y|X)) = \max_{p(x_i)} (H(Y)) - H(Y|X)$
  - Should maximize $H(Y)$
- If channel is also uniform with respect to the output:
  - same values on columns of P(Y|X)
  - $p(y_j) = \sum_i p(y_j|x_i)p(x_i)$
  - if $p(x_i) =$ uniform $= \frac{1}{n}$, then $p(y_j) = \frac{1}{n} \sum_i p(y_j|x_i) =$ uniform
  - therefore $p(y_j)$ are constant = uniform = H(Y) is maximized
  - H(Y) is maximized when H(X) is maximized (equiprobable symbols)

# Computing the capacity

- If channel is symmetric: use both tricks
  - $C = \max_{p(x_i)} (H(Y)) - H(Y|X)$
  - H(Y) is maximized when H(X) is maximized (equiprobable symbols)

# Examples of channels and their capacity

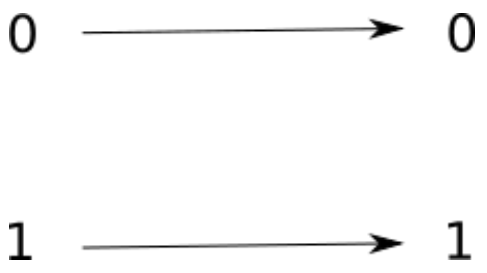



Figure 4: Noiseless binary channel

- Capacity $= 1$ bit/message, when $p(x_1) = p(x_2) = \frac{1}{2}$

# Noisy binary non-overlapping channel

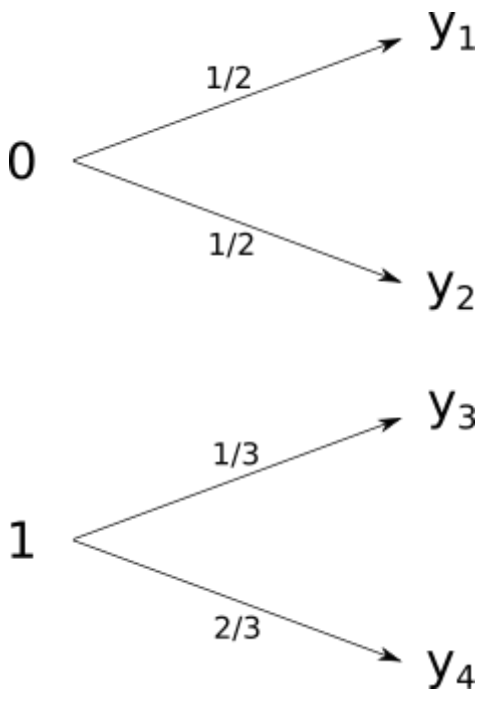



Figure 5: Noisy binary non-overlapping

- There is noise ($H(Y|X) > 0$), but can deduce the input ($H(X|Y) = 0$)
- Capacity = 1 bit/message, when $p(x_1) = p(x_2) = \frac{1}{2}$

# Noisy typewriter

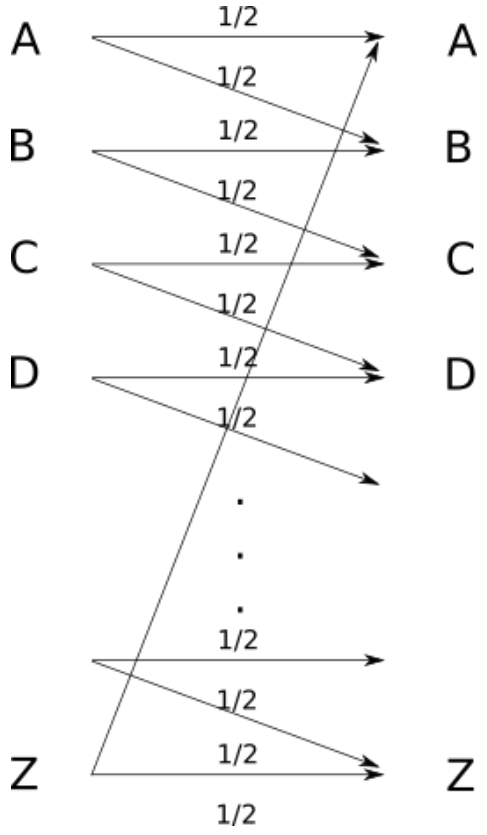



Figure 6: Noisy typewriter

$$\max I(X, Y) = \max (H(Y) - H(Y|X)) = \max H(Y) - 1$$
$$= \log(26) - 1 = \log(13)$$

# Noisy typewriter

- Capacity $=$ log(13) bit/message, when input probabilities are uniform
- Can transmit 13 letters with no errors (A, C, E, G, . . . )

# Binary symmetric channel

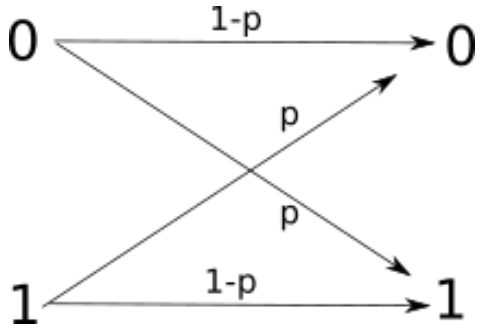



Figure 7: Binary symmetric channel (BSC)

- Capacity $= 1 - H_p = 1 + p\log(p) + (1-p)\log(1-p)$
- Capacity is reached when input distribution is uniform

# Binary erasure channel

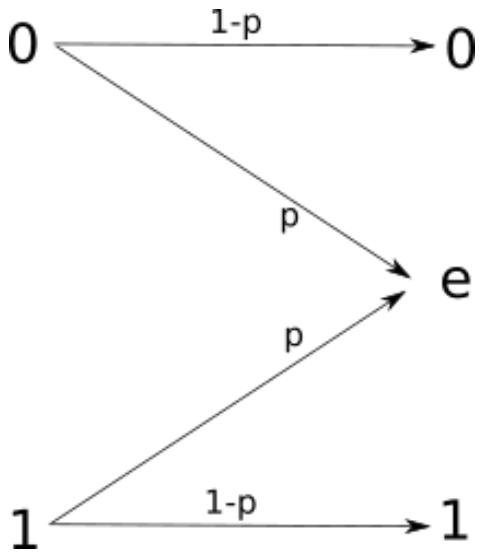



Figure 8: Binary erasure channel

- Different from BSC: here we know when errors happened
- Capacity $= 1 - p$
- Intuitive meaning: lose $p$ bits, remaining bits = capacity $= 1 - p$

# Symmetric channel of $n$-th order

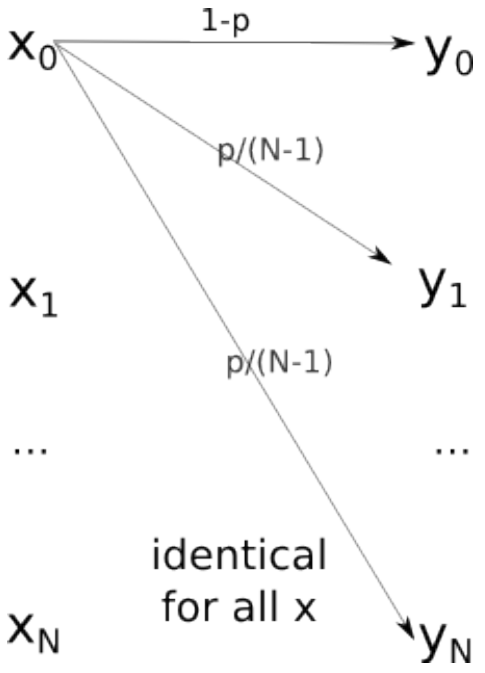



Figure 9: $N$-th order symmetric channel

- Extension of binary symmetric channel for $n$ symbols
- $1 - p$ chances that symbol has no error
- $p$ chances that symbol is changed, uniformly to any other (N-1) symbols ($\frac{p}{N-1}$ each)

# Symmetric channel of $n$-th order

- Channel is symmetric =>

$$C = \max_{p(x_i)} I(X,Y) = \max_{p(x_i)} (H(Y) - H(Y|X)) = \max_{p(x_i)} (H(Y)) - H(Y|X)$$

- $\max_{p(x_i)} (H(Y)) = \log(N)$
- $H(Y|X) = H(Y|x_i) =$ entropy of any row (same values)

==>

$$C = \log(N) + (1-p)\log(1-p) + p\log(\frac{p}{N-1})$$

- Capacity is reached when input probabilities are uniform

# Chapter summary

- Channel = Probabilistic system with two random variables $X$ and $Y$
- Characterization of transmission:
  - P(X,Y) => H(X,Y) *joint entropy*
  - $p(x_i)$, $p(y_j)$ *marginal distributions* => H(X), H(Y)
  - P(Y|X) *channel matrix* => H(Y|X) *average noise*
  - P(X|Y) => H(X|Y) *equivocation*
  - I(X,Y) *mutual information*
- Channel capacity: $C = \max_{p(x_i)} \ I(X, Y)$
- Examples:
  - Binary symmetric channel: $C = 1 - H_p$
  - Binary erasure channel: $C = 1 - p$
  - $N$-th symmetric channel: $C = \log(N) - H(of\ a\ row\ of\ channel\ matrix)$

# History

Figure 10: Claude Shannon (1916 - 2001)

- *A mathematical theory of communications*, 1948

# Exercises and problems

- At blackboard only