def: Let X, Y be R.V. The MUTUAL INFORMATION of X and Y is defined as:

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

$H(Y|X)$ is the conditional entropy of Y given X:

$$H(Y|X) = \sum_{x \in \mathcal{X}} \mathbb{P}[X=x] \cdot H(Y|X=x) = \sum_{x \in \mathcal{X}} \left[\mathbb{P}[X=x] \sum_{y \in \mathcal{Y}} \mathbb{P}[Y=y|X=x] \lg \frac{1}{\mathbb{P}[Y=y|X=x]} \right].$$

The mutual information is
- symmetric.
- a measure of correlation between X and Y.
- always ≥ 0.

→ the higher the mutual information, the more X and Y provide information about each other

→ the M.I between X and Y is minimal ($=0$) when X and Y are independant (since $H(Y|X) = H(Y)$).

We have seen that we cannot compress messages more than their entropy. Nevertheless, in real life, we hardly reach entropy when we communicate. This happens because the communication channel is NOISY, and having some redundancy guarantees more ROBUST communication.

→ On one hand, we want to transmit the least number of bits, meaning that we want to compress the information as much as possible (occupying the channel costs, faster communication). At the same time, we want the communication to be reliable, which can be achieved if we add some redundancy in the encoding.

→ The channel allows the communication but at the same time it introduces noise.

Shannon formulated a mathematical model to decode the source of a message while limiting communication errors.

Discrete Memoryless channels (DMC)

High level interpretation: the sender sends a sequence of symbols of a finite alphabet $\mathcal{X}$, $|\mathcal{X}| < \infty$.

The receiver receives a sequence of symbols, possibly from a different finite alphabet $\mathcal{Y}$, $|\mathcal{Y}| < \infty$.

Since the channel is noisy, some symbols can be altered and the receiver receives a different sequence.

MODEL:

Let $\underline{x} = x_1 \dots x_n$ be the message sent by the transmitter. In the channel there are $|\mathcal{X}|$ dices, one per symbol. Each dice has $|\mathcal{Y}|$ sides, each with different probabilities.

The probabilities are: $W(y|x) = \mathbb{P}[\text{receive } y \mid \text{sent } x]$.

The probabilities of each symbol in $\mathcal{X}$ and of each transmission are independent of one another, meaning that:

$$W^n(\underline{y}|\underline{x}) = \prod_{i=1}^{n} W(y_i|x_i) \qquad [2]$$

We can define a stochastic matrix $W(Y|X)$ as:

$$W(Y|X) = \begin{pmatrix} W(y_1|x_1) & \dots & W(y_{|\mathcal{Y}|}|x_1) \\ \vdots & & \\ W(y_1|x_{|\mathcal{X}|}) & \dots & W(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|}) \end{pmatrix} \qquad [3]$$

def: A DMC is a channel composed of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, $|\mathcal{X}|, |\mathcal{Y}| < \infty$ and a transition stochastic matrix $W(Y|X)$ as the one in [3] that satisfies [2].

def: An ENCODER $C^n \subseteq \mathcal{X}^n$ is a set of distinct sequences of length n that the transmitter can communicate through a channel.

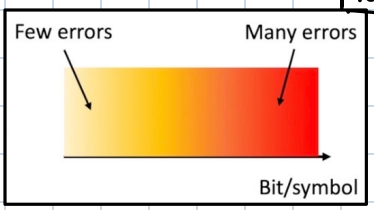
- it maps messages to "codewords". $|C^n|$ = # of different messages that can be transmitted through the channel.

def: A DECODER $\varphi_n : \mathcal{Y}^n \to C^n$ is a function that associates received messages of length n to sequences of C^n.

To be good, the communication must be s.t. the probability that a received message $\underline{y}$ is associated to the correct message $\underline{x}$ is high. At the same time, we want to compress the information as much as possible, so that we can send more information

↳ we want to maximise $\sum_{\underline{y} \in \varphi_n^{-1}(\underline{x})} W^n(\underline{y}|\underline{x})$ (= probability of receiving the correct messages) and $|C^n|$.

The problem is that these two quantities have inverse trends!



If each sent symbol carries a lot of information (expressed in bits), meaning that the information is strongly compressed, then an error occurring on the channel can cause the misunderstanding of the message.

Shannon's theorem, stated later, establishes the relationship between how much data can be sent on a MDC with an almost-zero error probability.

! GOAL = maximise the speed of the transmission (i.e. the size of C^n that can be transmitted in a unit of time), while keeping very low errors.
- HOW CAN WE DO IT? We CANNOT change W, that is an intrinsic characteristics of the channel. The only element that we can tune is C^n through $P(x)$

def: The transmission rate of an encoder C^n is the number of bits per transmission

$\frac{1}{n} \lg_2 |C^n|$ ↦ it is the number of bits carried by a symbol.

An error occurs when the transmitter sends a sequence $\underline{x}$, and the receiver receives a sequence $\underline{y}$ s.t. $\varphi(\underline{y}) \neq \underline{x}$, or, equivalently, if $\underline{y} \notin Im_{\varphi}(\underline{x})$

example. $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{a,b\}$. $\varphi_1 : \mathcal{Y} \to \mathcal{X} : \varphi_1(a)=0, \varphi_2(b)=1$

$$W(y|x) = \begin{matrix} W(y|x=0) \\ W(y|x=1) \end{matrix} \begin{pmatrix} \underset{0.8}{W(y=a|x)} & \underset{0.2}{W(y=b|x)} \\ 0.1 & 0.9 \end{pmatrix}$$

CORRECT

WRONG

$\underline{x} = (x_1, \dots x_n)$, $x_i \in \mathcal{X}$, $\underline{y} = (y_1, \dots, y_n)$, $y_i \in \mathcal{Y}$.

$\varphi_n : \mathcal{Y}^n \to \mathcal{X}$ s.t. $\varphi_n(\underline{x}) = (\varphi_1(x_1), \dots, \varphi_1(x_n))$

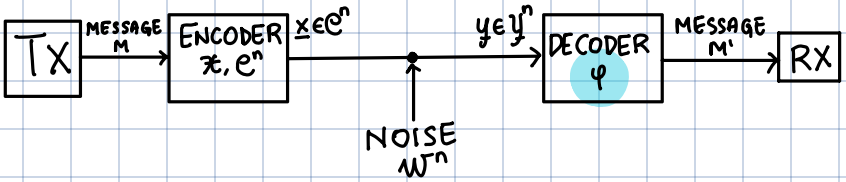
If $n=2$, then, $C^2 = \{00, 01, 10, 11\}$,

$\varphi_2(00) = aa$, $\varphi_2(01) = ab$, $\varphi_2(10) = ba$, $\varphi_2(11) = bb$.

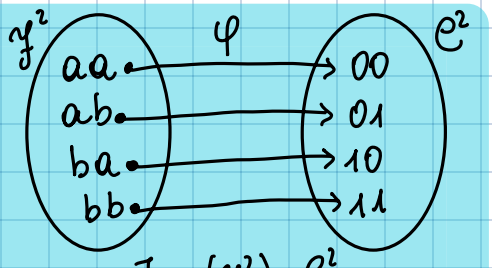
$P[aa|00] = W^2(0|0) = 0.8^2$. $P[ab|00] = 0.8 \cdot 0.2$

IF the sender sends $\underline{x} = 00$ and the receiver receives $\underline{y} = aa$, no errors occurred, because $aa \in Im^{-1}_{\varphi}(00)$

IF the sender sends $\underline{x} = 00$ and the receiver receives $\underline{y} = ab$, some error occurred, since $ab \notin Im^{-1}_{\varphi}(00)$



$Im_\varphi(\{aa\}) = \{00\}$, $Im^{-1}_\varphi(\{00\}) = \{aa\}$
$Im_\varphi(\{ab\}) = \{01\}$, $Im^{-1}_\varphi(\{01\}) = \{ab\}$
$Im_\varphi(\{ba\}) = \{10\}$, $Im^{-1}_\varphi(\{10\}) = \{ba\}$
$Im_\varphi(\{bb\}) = \{11\}$ $Im^{-1}_\varphi(\{11\}) = \{bb\}$



$Im_\varphi(\mathcal{Y}^2) = \mathcal{C}^2$
$Im^{-1}_\varphi(\mathcal{C}^2) = \mathcal{Y}^2$

def: the error probability for $\mathcal{C}^n$ and φ^n is:

$$W^n\left(\overline{\varphi_n^{-1}} \mid \underline{x}\right) = 1 - W^n\left(\varphi_n(\underline{x}) \mid \underline{x}\right)$$

this means $\mathcal{Y}^n - Im^{-1}_\varphi(\underline{x})$ [difference between sets]

In the previous example, if $\underline{x} = 00$, $\mathcal{Y}^2 - Im^{-1}_\varphi(00) = \{ab, ba, bb\}$.

$$W^n\left(\overline{\varphi_2^{-1}} \mid \underline{x}\right) = 1 - W(a|0)^2 = 1 - 0.8^2 = 0.36$$

def: the maximum error probability is:

$$e_n(W^n, \mathcal{C}^n, \varphi^n) = \max_{\underline{x} \in \mathcal{C}^n} W^n\left(\overline{\varphi_n^{-1}(\underline{x})} \mid \underline{x}\right)$$

def: $R \in \mathbb{R}$ is an ACHIEVABLE TRANSMISSION RATE for the DMC $(W^n, \mathcal{X}, \mathcal{Y})$ if $\exists$ sequence $\{\mathcal{C}^n, \varphi^n\}_{n \in \mathbb{N}}$ s.t.

$$e_n(W^n, \mathcal{C}^n, \varphi^n) \to 0 \qquad \overline{\lim_{n \to \infty}} \frac{1}{n} \lg |\mathcal{C}^n| \geq R$$

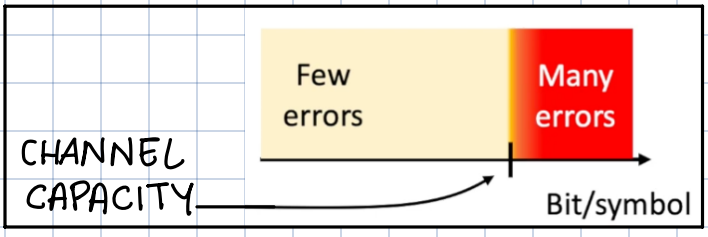
An achivable transmission rate is the number of bits that can be transmitted in a channel with an error that goes to zero.

OBS: all achievable rates are FINITE, $R < \infty$. This means that $\exists$ maximum achievable rate.

def: the capacity of a DMC W is the maximum achievable rate:

$$C(W) = \max_R \{R : R \text{ is an achievable rate}\}$$

So, basically, the capacity of a channel is the MAXIMUM ACHIEVABLE RATE (NUMBER OF BITS) THAT CAN BE TRANSMITTED WITH AN ERROR PROBABILITY THAT GOES TO ZERO. IT IS THE MAXIMUM RELIABLE TRASMISSION RATE.



def: The Shannon's capacity of a channel W is:

$$C^*(W) = \max_{(X,Y) \sim W(Y|X)} I(X,Y)$$

that is, the maximal mutual information between sent messages and received messages.

SHANNON'S CAPACITY THEOREM: $\boxed{C(W) = C^*(W)}$

This result is very powerful. It says that the capacity of a channel depends on how much information can be transmitted reliably over a noisy channel.
The model that we have seen decouples the communication process into two tasks:

- SOURCE CODING (data compression) → Reduce # of bits to represent data
- CHANNEL CODING (error correction) → Add redundancy to make transmission reliable.

Continuous channels

We've spoken about DISCRETE CHANNELS.

In IoT, we usually deal with CONTINOUS, WIRELESS CHANNELS

→ The Shannon capacity theorem has an alternative formulation in this case.

We consider a channel with Gaussian noise and a given bandwidth, B. The noise is independent of the signal.

- Gaussian noise.
  → we will see that channels are subject to ADDITIVE INTERFERENCES (SPOILER: WIRELESS communication happens through electromagnetic waves) that result in adding noise to the signal. Such noise usually follows a Gaussian or NORMAL distribution, which has probability density function

$$\varphi(z) = \frac{e^{-\frac{(z-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

where z is the noise value, μ is the mean and σ is the standard deviation

IF $\underline{x}$ is the input signal and y is the output signal, then we have:

$$y = \underline{x} + \underline{z}, \qquad \underline{z} \sim \mathcal{N}(0, \sigma_z)$$

→ We need to find $C^*(W) = \max_{(Y,X):\ P_{Y|X}=W} I(X,Y)$

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

$H(Y|X) = H(X+Z|X)$ but since $X \perp\!\!\!\perp Z$, $H(X+Z|X) = H(Z)$

Hence $I(X,Y) = H(Y) - H(Z)$.

What is the $P_{Y|X}$ s.t. $I(X,Y)$ is maximal?

→ When also $X \sim N(0, \sigma_X)$! [proof omitted].

There is a theorem in probability that says that if X, Z are normally distributed R.V. and are independent, and have variance σ_X^2 and σ_Z^2 respectively, then: $X+Z$ is also normally distributed and its variance is $\sigma_{X+Z}^2 = \sigma_X^2 + \sigma_Z^2$

The problem is that we have always spoken about discrete random variables, and we have defined entropy for discrete probability distributions.

Good news is that we can define entropy for continous R.V.s too! [differential entropy]

Without digging into details, the entropy for a normally distributed R.V. with variance σ is

$$\boxed{\frac{1}{2} \lg (2\pi e \sigma^2)}$$

Euler constant

So: $$C^*(W) = \frac{1}{2}\lg(2\pi e \sigma_Y^2) - \frac{1}{2}\lg(2\pi e \sigma_Z^2) =$$

$$= \frac{1}{2}\lg\left(\frac{2\pi e \sigma_Y^2}{2\pi e \sigma_Z^2}\right) = \frac{1}{2}\lg\left(\frac{\sigma_Y^2}{\sigma_Z^2}\right) = \frac{1}{2}\lg\left(\frac{\sigma_X^2+\sigma_Z^2}{\sigma_Z^2}\right) =$$

$$= \frac{1}{2}\lg\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right) = \left[\begin{array}{l}\text{The VARIANCE of a signal is equal to its POWER. THE QUANTITY} \\ \frac{\sigma_X^2}{\sigma_Z^2} \text{ is the SIGNAL TO NOISE RATIO (SNR)}\end{array}\right.$$

$$= \frac{1}{2}\lg(1+SNR)$$

So far, we have considered only the transmission of one signal $\underline{x}$. But we can send a sequence of signals. How long the sequence of signals can be is determined by the bandwidth of the channel, B. Hence we get the

SHANNON-HARTLEY THEOREM $\quad C^*(W) = B \lg(1+SNR)$

(Follows from the NYQUIST-SHANNON theorem. We have mentioned it in our 4th lecture - EMBEDDED SYSTEMS).