Chapter 5

Channel Capacity and Coding

In the previous chapters we faced with the problem of source coding. Once encoded, the information must be transmitted through a *communication channel* to reach its destination. This chapter is devoted to the study of this second step of the communication process.

5.1 Discrete Memoryless Channel

Each communication channel is characterized by the relation between the input and the output. For simplicity, throughout the analysis, we consider only discrete time channels. We know that, from an information theory perspective, the signals carry information and then they have a random nature; specifically they are stochastic processes $x(k,t)$. According to Shannon's sampling theorem, which also holds for random signals, if the signal bandwidth is limited, we can consider its samples[1] and then we can assume that the channel is discrete in time. The sampling of the stochastic process yields at the input of the channel the sequence of random variables $x(k,nT)$, as depicted in Figure 5.1. To ease the notation, we refer to the sequence $x(k,nT)$ as a sequence of random variables X_n, omitting the dependence on k. Clearly, the channel input can be seen as the outcome of an information source.
As to the values assumed by each random variable X_n, if the input source has a finite alphabet ($|\mathcal{X}| < \infty$) we have a discrete channel, a continuous channel otherwise.

[1] As a matter of fact the requirement of limited bandwidth is not necessary due to the presence of the channel which acts itself as bandwidth limiter.

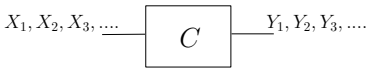

Figure 5.1: Discrete time channel. The input sequence is the sampling the stochastic process $x(k,t)$ with sampling step T.

5.1.1 A Mathematical Model for the channel

There are many factors, several of which with a random nature, that in a physical channel cause the output to be different from the input, e.g. attenuation, multipath, noise. Then, the input-output relation in a communication channel is, generally, a stochastic relation.

Definition. A *discrete channel* is a statistical model with an input X_n and an output Y_n which can be seen as a *noisy* version of X_n. The sequences X_n and Y_n take value in $\mathcal{X}$ and $\mathcal{Y}$ respectively ($|\mathcal{X}|, |\mathcal{Y}| < \infty$).

Given the input and the output alphabet $\mathcal{X}$ and $\mathcal{Y}$, a *channel* is described by the probabilistic relationship between the input and the output, i.e. by the set of *transition probabilities*

$$Pr\{Y_k = y | X_1 = x_1, X_2 = x_2, ..., X_k = x_k\} \quad y \in \mathcal{Y}, (x_1, ..., x_k) \in \mathcal{X}^k \quad (5.1)$$

where k denotes the discrete time at which the outcome is observed. Note that, due to causality, conditioning is restricted to the inputs preceding k and to the k-th input itself.
The channel is said *memoryless* when the output symbol at a given time depends only on the current input. In this case the transition probabilities become:

$$Pr\{Y_k = y | X_k = x\} \quad \forall y \in \mathcal{Y}, \forall x \in \mathcal{X}. \quad (5.2)$$

and the simplified channel scheme is illustrated in Figure 5.2. Assuming a memoryless channel greatly restricts our model since in this way we do not consider several factors, like fading, which could affect the communication because due to the introduction of intersymbol interference. Such phenomena require the adoption of much more complex models.
In order to further simplify the analysis, we also assume that the channel is *stationary*. Frequently[2], we can make this assumption without loss of generality since the channel variability is slow with respect to the transmission rate. In other words, during the transmission of a symbol, the statistical properties of the channel do not change significantly. Then, since the probabilistic

[2]This is not true when dealing with mobile channels.

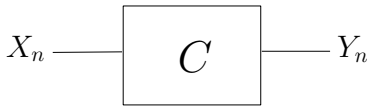

Figure 5.2: Discrete memoryless channel. The output signal at each time instant n (r.v.) depends on the input signal (r.v.) at the same time.

model describing the channel does not change over time, we can characterize the channel by means of the transition probabilities $p(y|x)$, where $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. These probabilities can be conveniently arranged in a matrix $\mathbf{P} = \{P_{ij}\}$, where

$$P_{ij} = P\{y_j|x_i\} \quad j = 1, .., |\mathcal{Y}| \quad i = 1, ..., |\mathcal{X}|. \tag{5.3}$$

The matrix $\mathbf{P}$ is called *channel matrix* or *channel transition matrix*.

5.1.2 Examples of discrete memoryless channels

Noiseless binary channel

Suppose that we have a channel in which the binary input is reproduced exactly at the output. Then, any transmitted bit is received without error. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.4}$$

This is a limit case, for which we have no longer a probabilistic channel. A graphical representation of the noiseless channel is given in Figure 5.3.

Noisy channel with non-overlapping outputs

This is another example in which noise does not affect the transmission, even if the channel is probabilistic. Indeed, see Figure 5.4, the output of the channel depends randomly on the input; however the input can be exactly determined from the output and then every transmitted bit can be recovered without any error. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}. \tag{5.5}$$

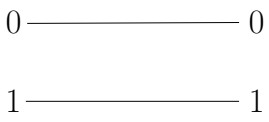

Figure 5.3: Noiseless binary channel.

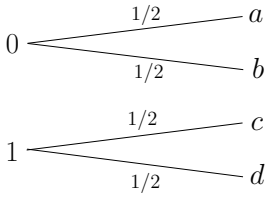

Figure 5.4: Model of the noisy channel with non overlapping outputs.

Noisy Typewriter

This is a more realistic example. A channel input is delivered unchanged at the output with probability 1/2 and transformed into the subsequent element with probability 1/2. In this case, the transmitted signal can not be correctly recovered from the output. Figure 5.5 illustrates the behavior of this channel; the transition matrix has the following form

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & ... & 0 \\ 0 & 1/2 & 1/2 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & 0 \\ 1/2 & 0 & ... & ... & ... & 1/2 \end{pmatrix}. \tag{5.6}$$

Binary Symmetric Channel (BSC)

The binary symmetric channel is a binary channel in which the input symbols are flipped with probability ε and left unchanged with probability $1-\varepsilon$ (Figure 5.6). The transition matrix of the BSC is

$$\mathbf{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}. \tag{5.7}$$

This channel model is used very frequently in communication engineering. Without loss of generality, we will only consider BSC with $\varepsilon < 1/2$. Indeed, if $\varepsilon > 1/2$, we can trivially reverse the input symbols thus yielding an error probability lower than 1/2.

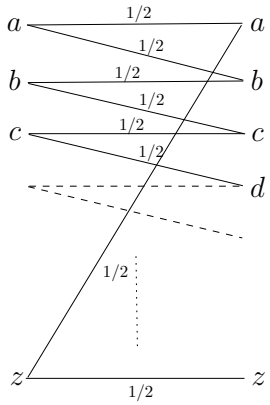

Figure 5.5: Noisy typewriter.

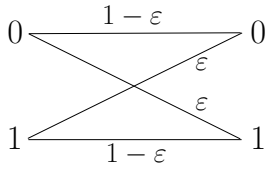

Figure 5.6: Binary symmetric channel.

Binary Erasure Channel (BEC)

This channel is similar to the binary symmetric channel, but in this case the bits are lost, rather than flipped, with a given probability α. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{pmatrix}. \tag{5.8}$$

The channel model is depicted in Figure 5.7.

5.2 Channel Coding

From previous chapters we know that $H(X)$ represents the fundamental limit on the rate at which a discrete memoryless source can be encoded. We we will prove that a similar fundamental limit also exists for the transmission rate over communication channels.
The main goal when transmitting information over any communication channel is *reliability*, which is measured by the probability of correct reception at the output of the channel. The surprising result that we will prove in this

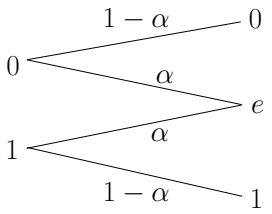

Figure 5.7: Binary erasure channel.

chapter is that reliable transmission is possible even over noisy channels, as long as the transmission rate is sufficiently low. The existence of a fundamental bound on the transmission rate, proved by Shannon, is one of the most remarkable results of information theory.
By referring to the example of the noisy typewriter in Section 5.1.2, some interesting considerations can be made. By using only half of the inputs, it is possible to make the corresponding outputs disjoint, and then recover the input symbols from the output. Then, this subset of the inputs can be transmitted over the channel with no error. This is just an example in which the limitation that the noise causes in the communication is not on the reliability of the communication but on the rate of the communication. This example provides also a first insight into channel coding: limiting the inputs to a subset is similar to the addition of redundancy which will be performed through channel coding.

5.2.1 Preview of the Channel Coding Theorem

BSC: a qualitative analysis

By looking at the binary symmetric channel we try to apply a similar approach to that used for the noisy typewriter in order to determine if non-overlapping outputs, and then transmission without error, can be obtained in the BSC case. To this purpose, we have to consider sequences of input symbols instead of single inputs. Then, we define the n-th extension of the channel or *extended channel*, which is a channel having input and output alphabets $\mathcal{X}^n = \{0,1\}^n$ and $\mathcal{Y}^n = \{0,1\}^n$ and transition probabilities $p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$. Figure 5.8 gives a schematic representation of the extended channel. Due to the dispersion introduced by the channel, a set of possible output sequences corresponds to a n-length transmitted sequences. If the sets corresponding to different input sequences were disjoint, the transmission would be error-free. This happens only with channels having non-overlapping outputs. By looking at the BSC, Figure 5.6, we see that it is no so, but we can consider a subset of the input sequences in order

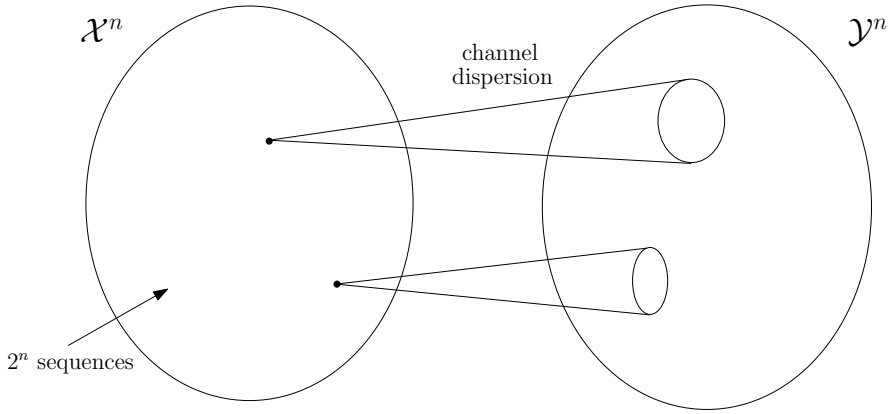

Figure 5.8: Representation of the n-th extension of the channel.

to make the corresponding set disjoint. That is, we can consider 2^k input sequences for some value k ($k < n$). Note that, without noise, k bits would suffice to index 2^k sequences; the $n - k$ additional bits in each sequence correspond to the 'redundancy'. In the sequel we better formalize this concept.

In the BSC, according to the law of large numbers, if a binary sequence of length n (for large n) is transmitted over the channel with high probability, the output will disagree with the input at about $n\varepsilon$ positions. The number of possible ways in which it is possible to have $n\varepsilon$ error in a n-length sequence (or the number of possible sequences that disagree with the input in $n\varepsilon$ positions) is given by

$$\binom{n}{n\varepsilon}. \tag{5.9}$$

By using Stirling's approximation $n! \approx n^n e^{-n}\sqrt{2\pi n}$ and by applying some algebra we obtain

$$\binom{n}{n\varepsilon} \approx \frac{2^{nh(\varepsilon)}}{\sqrt{2\pi n(1-\varepsilon)\varepsilon}}. \tag{5.10}$$

Relation (5.10) gives an approximation on the number of sequences in each output set. Then, for each block of n inputs, there exist roughly $2^{nh(\varepsilon)}$ highly probable corresponding output blocks. Note that if $\varepsilon = 1/2$, then $h(\varepsilon) = 1$ and the entire output set would be required for an error-free transmission of only one input sequence.

On the other hand, by referring to the output of the channel, regarded as a source, the total number of highly probable sequences is roughly $2^{nH(Y)}$. Therefore, the maximum number of input sequences that may produce almost

non-overlapping output sets is at most equal to

$$M = \frac{2^{nH(Y)}}{2^{nh(\varepsilon)}/\sqrt{2\pi n(1-\varepsilon)\varepsilon}}. \tag{5.11}$$

As a consequence, the maximum number of *information bits* that can be correctly transmitted is

$$k = \log_2 \left(2^{n(H(Y)-h(\varepsilon))} \cdot \sqrt{2\pi n(1-\varepsilon)\varepsilon} \right). \tag{5.12}$$

Then, the number of bit that can be transmitted each time, i.e. the transmission rate for channel use is:

$$R = \frac{k}{n} = \frac{\log_2(2^{n(H(Y)-h(\varepsilon))} \cdot \sqrt{2\pi n(1-\varepsilon)\varepsilon})}{n}. \tag{5.13}$$

Finally, as $n \to \infty$, $R \to H(Y) - h(\varepsilon)$.
A close inspection of the limit expression for R reveals that we have still a degree of freedom that can be exploited to maximize the transmission rate; it consists in the input probabilities $p(x)$, which determine the values of $p(y)$ (remember that the transition probability of the channel are fixed by the stationarity assumption) and then $H(Y)$. In the sequel we look for the input probability distribution maximizing $H(Y)$, giving the maximum transmission rate. Since Y is a binary source, the maximum of $H(Y)$ is 1, which is obtained when the input symbols are equally likely. So, the maximum transmission rate is $R_{max} = 1 - h(\varepsilon)$.

Observation.
The quantity $1-h(\varepsilon)$ is exactly the maximum value of the mutual information between the input and the output for the binary symmetric channel (BSC), that is

$$\max_{p_X(x)} I(X;Y) = 1 - h(\varepsilon). \tag{5.14}$$

In fact, given the input bit x, the BSC behaves as a binary source, giving at the output the same bit with probability $1 - \varepsilon$. Thus, we can state that $H(Y|X) = h(\varepsilon)$ and consequently $I(X;Y) = H(Y) - H(Y|X) = H(Y) - h(\varepsilon)$, whose maximum is indeed $1 - h(\varepsilon)$.

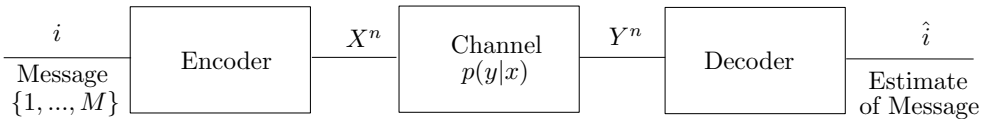

Figure 5.9: Communication channel.

Qualitative analysis of a general discrete memoryless channel

The previous analysis explains the essence of Shannon's theorem on the channel coding by focusing specifically on the binary symmetric channel. In order to extend the previous analysis to a generic channel we need some clarifications. Firstly, we note that when we refer to sets of outputs we do not mean necessarily a compact set. Given an input, the corresponding output sequence may be scattered throughout the whole space $\mathcal{Y}^n$, depending on the behavior of the channel. Secondly, the output sets in a general channel have usually different sizes since the channel is not symmetric.
We can affirm that, given an input sequence x^n, the number of possible output sequences y^n is approximately $2^{nH(Y|x^n)}$, with high probability. This is indeed the approximate number of typical sequences with respect to the distribution $p(y|X^n = x^n)$. By varying the input sequence x^n, we can consider the mean number of output sequences $2^{nH(Y|X)}$. Since the total number of typical sequences for the source Y is still $2^{nH(Y)}$, it follows that the maximum number of disjoint sets is $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Accordingly, we can correctly transmit $I(X;Y)$ information bits for channel use. By properly choosing the prior probabilities, we directly have the following expression for the maximum achievable rate:

$$R_{max} = \max_{p_X(x)} I(X;Y). \tag{5.15}$$

This result is in agreement with the previous one for the BSC. We foretell that the above expression represents the channel capacity.
In Section 5.2.3, we will give a rigorous formalization to the above considerations by proving the *noisy channel-coding theorem*.

5.2.2 Definitions and concepts

Let $\{1, 2, ..., M\}$ be the index set from which a message is drawn. Before being transmitted into the channel the indexes are encoded. At the receiver side, by observing the output of the channel the receiver guesses the index through an appropriate decoding rule. The situation is depicted in Figure 5.9. Let us rigorously define some useful concepts, many of them already

discussed in the previous section.

Definition. A *discrete memoryless channel* (DMC) consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability mass functions $p(y|x)$, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$.

Definition. The *nth extension of the discrete memoryless channel* corresponds to the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \quad k = 1, 2, ..., n \tag{5.16}$$

i.e. the output does not depend on the past inputs and outputs.
If the channel is used *without feedback*, i.e. if the input symbols do not depend on the past output symbols ($p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$), the channel transition probabilities for the nth extension of the DMC can be written as

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i). \tag{5.17}$$

We shall always implicitly refer to channels without feedback, unless stated otherwise.

Definition. An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of:
1. An encoding function $g : \{1 : M\} \to \mathcal{X}^n$, which is a mapping from the index set to a set of codewords or *codebook*.
2. A decoding function $f : \mathcal{Y}^n \to \{1 : M\}$, which is a deterministic rule assigning a number (index) to each received vector.

Definition. Let λ_i be the error probability given that index i was sent, namely the *conditional probability of error*:

$$\lambda_i = Pr\{f(y^n) \neq i | x^n = g(i)\}. \tag{5.18}$$

Often, we will use $x^n(i)$ instead of $g(i)$ to indicate the codeword associated to index i. As a consequence of the above definition, the *maximal probability of error* $\lambda_{max}^{(n)}$ for an (M, n) code is defined as

$$\lambda_{max}^{(n)} = \max_{i \in \{1,2,...,M\}} \lambda_i. \tag{5.19}$$

The *average probability of error* $P_e^{(n)}$ for an $(M.n)$ code is

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i, \tag{5.20}$$

where we implicitly assumed that the indexes are drawn in an equiprobable manner. We point out that the average probability of error, like the maximum one, refers to the n-length sequences.

Definition. The *rate* R of an (M, n) code is

$$R = \frac{\log M}{n} \quad \text{bits per channel use.} \tag{5.21}$$

Definition. A rate R is said to be *achievable* if there exists a sequence of codes having rate R, i.e. $(2^{nR}, n)$ codes, such that

$$\lim_{n\to\infty} \lambda_{max}^{(n)} = 0. \tag{5.22}$$

Definition. The *capacity* of the channel is the supremum of all the achievable rates.

Jointly typical sequences and set

In order to describe the decoding process in Shannon's coding theorem it is necessary to introduce the concept of 'joint typicality'.

Definition. Given two DMSs X and Y, the set $A_\varepsilon^{(n)}$ of *joint typical* sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$, is the following set of n-long sequences

$$\begin{aligned} A_\varepsilon^{(n)} = \{&(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ &\left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \varepsilon, \left|-\frac{1}{n}\log p(y^n) - H(Y)\right| < \varepsilon, \\ &\left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \varepsilon\Big\}, \end{aligned} \tag{5.23}$$

where the first and the second conditions require the typicality of the sequences x^n and y^n respectively, and the last inequality requires the joint typicality of the couple of sequences (x^n, y^n).

We observe that if we do not considered the joint typicality, the number of possible couples of sequences in $A_\varepsilon^{(n)}$ would be the product $|A_{\varepsilon,x}^{(n)}| \cdot |A_{\varepsilon,y}^{(n)}| \cong 2^{n[H(X)+H(Y)]}$. The intuition suggests that the total number of jointly typical sequences is approximately $2^{nH(X,Y)}$ and then not all pairs of typical x^n and typical y^n are jointly typical since $H(X,Y) \leq H(X) + H(Y)$. These considerations are formalized in the following theorem, which is the extension of the AEP theorem to the case of two sources.

Theorem (*joint AEP*)**.**
Let X and Y be two DMS with marginal probabilities p_X and p_Y and let (x^n, y^n) be a couple of sequences of length n drawn from the two sources. Then:

1. $\Pr\{A_\varepsilon^{(n)}\} \to 1 \quad \text{as} \quad n \to \infty \quad (> 1-\delta \quad \text{for large } n)$;

2. $\forall \varepsilon, \quad |A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)} \quad \forall n$;

3. $\forall \delta, \forall \varepsilon, n \text{ large}, \quad |A_\varepsilon^{(n)}| \geq (1-\delta)2^{n(H(X,Y)-\varepsilon)}$;

4. Considering two sources $\tilde{X}$ and $\tilde{Y}$ with alphabets $\mathcal{X}$ and $\mathcal{Y}$ such that $p_{\tilde{X}} = p_X$ and $p_{\tilde{Y}} = p_Y$ but independent of each other, i.e. such that $(\tilde{X}^n, \tilde{Y}^n) \sim p_X(x^n)p_Y(y^n)$, we have

$$Pr\{(\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}\} \cong 2^{-nI(X;Y)}. \tag{5.24}$$

Formally,

$$\forall \varepsilon < 0, \forall n, \quad Pr\{(\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}\} \leq 2^{-n(I(X;Y)-3\varepsilon)} \quad , \tag{5.25}$$

and

$$\forall \varepsilon > 0, \forall \delta > 0, n \text{ large}, \quad Pr\{(\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}\} \geq (1-\delta)2^{-n(I(X;Y)+3\varepsilon)}. \tag{5.26}$$

Proof. The first point says that for large enough n, with high probability, the couple of sequences (x^n, y^n) lies in the typical set. It directly follows from the weak law of large numbers. In order to prove the second and the third point we can use the same arguments of the proof of the AEP theorem. Instead, we explicitly give the proof of point 4 which represents the novelty with respect to the AEP theorem. The new sources $\tilde{X}^n$ and $\tilde{Y}^n$ are independent but have

the same marginals as X^n and Y^n, then

$$\begin{aligned} Pr\{(\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}\} &= \sum_{(\tilde{x}^n,\tilde{y}^n)\in A_\varepsilon^{(n)}} p_{\tilde{X}}(\tilde{x}^n) p_{\tilde{Y}}(\tilde{y}^n) \\ &= \sum_{(\tilde{x}^n,\tilde{y}^n)\in A_\varepsilon^{(n)}} p_X(\tilde{x}^n) p_Y(\tilde{y}^n) \\ &= \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p_X(x^n) p_Y(y^n) \\ &\overset{(a)}{\leq} |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)} \\ &\overset{(b)}{\leq} 2^{-n(H(X,Y)+\varepsilon)} 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)} \\ &= 2^{-n(I(X;Y)-3\varepsilon)}, \end{aligned} \tag{5.27}$$

where inequality (a) follows from the AEP theorem, while (b) derives from point 2. Similarly, it's possible to find a lower bound for sufficiently large n, i.e.

$$\begin{aligned} Pr\{(\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}\} &= \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n) p(y^n) \\ &\geq (1-\delta) 2^{-n(H(X)+H(Y)-H(X,Y)+3\varepsilon)} \\ &\geq (1-\delta) 2^{-n(I(X;Y)+3\varepsilon)}. \end{aligned} \tag{5.28}$$

□

The above theorem suggests that we have to consider about $2^{nI(X;Y)}$ pairs before we are likely to come across a jointly typical pair.

5.2.3 Channel Coding Theorem

We are now ready to prove the other basic theorem of information theory stated by Shannon in 1948, that is the *channel coding theorem*. As previously mentioned, the remarkable result of this theorem is that, even though the channel introduce errors, the information can still be reliably sent over the channel at all rates up to channel capacity. Shannon's key idea is to sequentially use the channel many times, so that the law of large number comes into effect. Shannon's outline of the proof is indeed strongly based on the concept of typical sequences and in particular on a joint typicality based

decoding rule. However, the rigorous proof was given long after Shannon's initial paper. We now give the complete statement and proof of Shannon's second theorem.

Theorem (*Channel Coding Theorem).*
Let us define the channel capacity as follows:

$$C = \max_{p_X(x)} I(X;Y). \tag{5.29}$$

For a discrete memoryless channel a rate R is achievable *if and only if* $R < C$.

According to the definition of achievable rate, the direct implication states that, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$. Conversely, the reverse implication says that for any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, $R \leq C$.
Let us now prove that all the rates $R < C$ are achievable (direct implication, *if*). Later we will prove that any rate exceeding C is not achievable (converse implication,*only if*).

Proof. (*Channel Coding Theorem: Achievability*)
Let us fix $p_X(x)$.
For any given rate R, we have to find a proper sequence of $(2^{nR}, n)$ codes. The question that arises is how to build a codebook. It may come as a surprise that Shannon suggests to take the codewords at random. Specifically, we generate a $(2^{nR}, n)$ code according to the distribution $p(x)$ by taking 2^{nR} codewords drawn according to the distribution $p(x^n) = \prod_{i=1}^{n} p(x_i)$, thus obtaining a mapping

$$g : \{1, 2, ..., 2^{nR}\} \to \mathcal{X}^n. \tag{5.30}$$

We can organize the codewords in a matrix $2^{nR} \times n$ as follows

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & \dots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{pmatrix}. \tag{5.31}$$

Each element of the matrix is drawn i.i.d. $\sim p(x)$. Each row i of the matrix corresponds to the codeword $x^n(i)$.
Having defined the encoding function g, we define the correspondent decoding function f. Shannon proposed a decoding rule based on *joint typicality*. The receiver looks for a codeword that is jointly typical with the received sequence. If a unique codeword exists satisfying this property, the receiver

declares that word to be the transmitted codeword. Formally, given y^n, if the receiver finds a unique i s.t. $(y^n, x^n(i)) \in A_\varepsilon^{(n)}$, then

$$f(y^n) = i. \tag{5.32}$$

Otherwise, that is if no such i exists or if there is more than one such codeword, an error is declared and the transmission fails. Notice that joint typical decoding is suboptimal. Indeed, the optimum procedure for minimizing the probability of error is the maximum likelihood decoding. However the proposed decoding rule is easier to analyze and asymptotically optimal.

We now calculate the average probability of error over all codes generated at random according to the above described procedure, that is

$$P_e^{(n)} = \sum_{\mathcal{C}} P_e^{(n)}(\mathcal{C}) Pr(\mathcal{C}) \tag{5.33}$$

where $P_e^{(n)}(\mathcal{C})$ is the probability of error averaged over all codewords in codebook $\mathcal{C}$. Then we have[3]

$$\begin{aligned} P_e^{(n)} &= \sum_{\mathcal{C}} Pr(\mathcal{C}) \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}) \\ &= \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_i(\mathcal{C}). \end{aligned} \tag{5.34}$$

By considering the specific code construction we adopted, it's easy to argue that λ_i does not depend on the particular index i sent. Thus, without loss of generality, we can assume $i = 1$, yielding

$$P_e^{(n)} = \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_1(\mathcal{C}). \tag{5.35}$$

If Y^n is the result of sending $X^n(i)$ over the channel[4], we define the event E_i as the event that the i-th codeword and the received one are jointly typical, that is

$$E_i = \{(X^n(i), Y^n) \in A_\varepsilon^{(n)}\}, \quad i \in \{1, 2, ..., 2^{nR}\}. \tag{5.36}$$

[3]We precise that there is a slight abuse of notation, since $P_e^{(n)}(\mathcal{C})$ in (5.33) corresponds to $P_e^{(n)}$ in (5.20), while $P_e^{(n)}$ in (5.33) denotes the probability of an error averaged over all the codes. Similarly, $\lambda_i(\mathcal{C})$ corresponds to λ_i where again the dependence on the codebook is made explicit.

[4]Both $X^n(i)$ and Y^n are random since we are not conditioning to a particular code. We are interested in the average on C.

Since we assumed $i = 1$, we can define the *error event* $\mathcal{E}$ as the union of all the possible types of error which may occur during the decoding procedure (jointly typical decoding):

$$\mathcal{E} = E_1^c \cup E_2 \cup E_3 \cup ... \cup E_{2^{nR}}, \tag{5.37}$$

where the event E_1^c occurs when the transmitted codeword and the received one are not jointly typical, while the other events refer to the possibility that a wrong codeword (different from the transmitted one) is jointly typical with Y^n (the received sequence). Hence: $Pr(\mathcal{E}) = Pr(E_1^c \cup E_2 \cup E_3 \cup ... \cup E_{2^{nR}})$. We notice that the transmitted codeword and the received sequence must be jointly typical, since they are probabilistically linked through the channel. Hence, by bounding the probability of the union in (5.37) with the sum of the probabilities, from the first and the fourth point of the joint AEP theorem we obtain

$$\begin{aligned} Pr(\mathcal{E}) &\leq Pr(E_1^c) + \sum_{i=2}^{2^{nR}} Pr(E_i) \\ &\leq \delta + \sum_{i=1}^{2^{nR}} 2^{-n(I(X;Y)-3\varepsilon)}, \\ &\leq \delta + (2^{nR} - 1)2^{-n(I(X;Y)-3\varepsilon)}, \\ &\leq \delta + 2^{nR}2^{-n(I(X;Y)-3\varepsilon)}, \\ &\leq \delta + 2^{-n(I(X;Y)-R-3\varepsilon)} \\ &= \delta', \end{aligned} \tag{5.38}$$

where δ' can be made arbitrarily small for $n \to \infty$ if $R < I(X;Y)$. The intuitive meaning of the above derivation is the following: since for any codeword, different from the transmitted one, the probability to be jointly typical with the received sequence is approximately $2^{-nI(X;Y)}$, we can use at most $2^{nI(X;Y)}$ codewords in order to keep the error probability arbitrarily small for large enough n. In other words, if we have not too many codewords ($R < I$), with high (arbitrarily close to 1) probability there is no other codeword that can be confused with the transmitted one.
At the beginning of the proof, we fixed $p_X(x)$ which determines the value of $I(X;Y)$. Actually $p_X(x)$ is the ultimate degree of freedom we can exploit in order to obtain the smallest $Pr(\mathcal{E})$ for the given rate R. As a consequence, it is easy to argue that $Pr(\mathcal{E})$ can be made arbitrarily small (for large n) if

the rate R is less than the maximum of mutual information, that is

$$C = \max_{p_X(x)} I(X;Y). \tag{5.39}$$

To conclude the proof we need a further step. In fact, the achievability definition is given in terms of the maximal probability of error $\lambda_{max}^{(n)}$, while up to now we have dealt with the average probability of error. We now show that

$$P_e^{(n)} \to 0 \quad \Rightarrow \quad \exists \mathcal{C} \quad \text{s.t} \quad \lambda_{max}^{(n)} \to 0. \tag{5.40}$$

Since $P_e^{(n)} = Pr(\mathcal{E}) < \delta'$, there exists at least one code $\mathcal{C}$ (actually more than one) such that $P_e^{(n)}(\mathcal{C}) < \delta'$. Name it $\mathcal{C}^*$. Let us list the probabilities of error λ_i of the code $\mathcal{C}^*$ in increasing order:

$$\lambda_1, \lambda_2, ..., \lambda_{2^{nR}}.$$

Now, we throw away the upper half of the codewords in $\mathcal{C}^*$, thus generating a new code $\mathcal{C}^\star$ with half codewords. Being the average probability of error for the code $\mathcal{C}^*$ lower than δ' we deduce that

$$\lambda_{\frac{2^{nR}}{2}} < 2\delta'. \tag{5.41}$$

(If it were not so, it is easy to argue that $P_e^{(n)}(\mathcal{C})$ would be greater than δ'.) But $\lambda_{2^{nR}/2}$ is the maximal probability of error for the code $\mathcal{C}^\star$, which then is arbitrarily small (tends to zero as $n \to \infty$).
What about the rate of $\mathcal{C}^\star$? Throwing out half the codewords reduces the rate from R to $R - \frac{1}{n}$ $(= \log(2^{nR-1})/n)$. This reduction is negligible for large n. Then, for large n, we have found a code having rate R and whose $\lambda_{max}^{(n)}$ tends to zero. This concludes the proof that any rate below C is achievable. □

Some considerations can be made regarding the proof: similarly to the source coding theorem, Shannon does not provide any usable way to construct the codes. The construction procedure used in the proof is highly impractical for many reasons. Firstly, Shannon's approach is asymptotical: both the number of codewords, 2^{nR}, and the length, n, have to go to infinity. Secondly, but not least, Shannon suggests to generate the code at random; accordingly, we should write down all the codewords in the matrix $\mathcal{C}$ (see (5.31)) and moreover transmit the matrix to the receiver. It is easy to guess that, for large values of n, this scheme requires (storage and transmission) resources out of any proportion. In fact, without some structure in the code

it is not possible to decode. Only structured codes (i.e. codes generated according to a rule) are easy to encode and decode in practice.

Now we must show that it is not possible to 'do better' than C (converse). Before giving the proof we need to introduce two lemmas of general validity.

Lemma (Fano's inequality)**.** *Let X and Y be two dependent sources and let g be any deterministic reconstruction function s.t. $\hat{X} = g(Y)$. The following upper bound on the remained uncertainty (or equivocation) about X given Y holds:*

$$\begin{aligned} H(X|Y) &\leq h(P_e) + P_e \log(|\mathcal{X}| - 1) \\ &\leq 1 + P_e \log(|\mathcal{X}| - 1), \end{aligned} \tag{5.42}$$

where $P_e = Pr(\hat{X} \neq X)$.

Proof. We introduce an error random variable

$$E = \begin{cases} 1 & \text{if } \hat{x} \neq x \quad \text{(with probability } P_e\text{)} \\ 0 & \text{if } \hat{x} = x \quad \text{(with probability } 1 - P_e\text{)}. \end{cases} \tag{5.43}$$

By using the chain rule we can expand $H(E, X|Y)$ in two different ways:

$$H(X, E|Y) = H(X|Y) + H(E|X, Y) \tag{5.44}$$

$$= H(E|Y) + H(X|E, Y). \tag{5.45}$$

It's easy to see that $H(E|X, Y) = 0$ while $H(E|Y) < H(E) = h(P_e)$. As to $H(X|E, Y)$, by expliciting the sum on E we have

$$H(X|E, Y) = (1 - P_e)H(X|0, Y) + P_e H(X|1, Y). \tag{5.46}$$

Relation (5.46) can be simplified by observing that, when $E = 0$, there is no uncertainty on the value of X (that is, being $\hat{x} = x$, $H(X|0, Y) = 0$) while, when $E = 1$, the estimation of X is not correct (being $\hat{x} \neq x$). Using the bound on the maximum entropy yields $H(X|1, Y) \leq \log(|\mathcal{X}| - 1)$. Then, the sum in (5.46) can be written as:

$$H(X|E, Y) \leq P_e \log(|\mathcal{X}| - 1). \tag{5.47}$$

By expliciting $H(X|Y)$ from equality (5.44)-(5.45) we eventually have

$$\begin{aligned} H(X|Y) &\leq h(P_e) + P_e \log(|\mathcal{X}| - 1) \\ &\leq 1 + P_e \log(|\mathcal{X}| - 1), \end{aligned} \tag{5.48}$$

which is the desired relation.
The second inequality provides a weaker upper bound which however allows to avoid the evaluation of the binary entropy $h(P_e)$. □

Fano's inequality is useful whenever we know a random variable Y and we wish to guess the value of a correlated random variable X. It relates the probability of error in guessing the random variable X, i.e. P_e, to the conditional entropy $H(X|Y)$.
It's interesting to note that Fano's inequality can also be seen as a lower bound on P_e. Looking at X and Y as the input and the output of a channel and looking at g as the decoding function, P_e corresponds to the probability of a decoding error[5].

Lemma. *Let us consider a discrete memoryless channel (DMC) with input and output sources X and Y. By referring to the extended channel we have*

$$I(X^n;Y^n) \leq nC. \tag{5.49}$$

Proof.

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &\overset{(a)}{=} H(Y^n) - \sum_i H(Y_i|Y_{i-1},...,Y_1,X^n) \\ &\overset{(b)}{=} H(Y^n) - \sum_i H(Y_i|X_i) \\ &\overset{(c)}{\leq} \sum_i H(Y_i) - \sum_i H(Y_i|X_i) \\ &= \sum_i I(Y_i;X_i) \leq nC, \end{aligned} \tag{5.50}$$

where (a) derives from the application of the generalized chain rule and (b) follows from the memoryless (and no feedback) assumption. Since conditioning reduces uncertainty, $H(Y^n) \leq \sum_i H(Y_i)$, we have relation (c). We stress that the output symbols Y_i do not need to be independent, that is generally $p(y_i|y_{i-1},...,y_1) \neq p(y_i)$. Since C is defined as the maximal mutual information over $p(x)$ the last inequality clearly holds. □

The above lemma shows that using the channel many times does not increase the transmission rate.

[5]For sake of clarity, we point out that Fano's inequality holds even in the more general case in which the function $g(Y)$ is random, that is for any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$.

Remark: the lemma holds also for non DM channels, but this extension is out of the scope of these notes.

We have now the necessary tools to prove the converse of the channel coding theorem.

Proof. (*Channel Coding Theorem: Converse*)
We show that any sequence of $(2^{nR},n)$ codes with $\lambda_{max}^{(n)} \to 0$ must have $R \leq C$; equivalently, if $R > C$ then $P_e^{(n)}$ cannot tend to 0 (thus implying that $\lambda_{max}^{(n)}$ does not tend to 0).
Given the index set $\{1, 2, ..., 2^{nR}\}$, a fixed encoding function which associates to a index (message) W a codeword $X^n(W)$, and a fixed decoding rule $g(\cdot)$ such that $\hat{W} = g(Y^n)$, we have

$$W \to X^n(W) \to Y^n \to \hat{W}. \tag{5.51}$$

In (5.51), Y^n takes the role of the observation, W the role of the index we have to estimate and $Pr(\hat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \sum_i \lambda_i$. The random variable W corresponds to a uniform source, since the indexes are drawn in an equiprobable manner, thus the entropy has the expression $H(W) = \log(2^{nR})$. By using the definition of the mutual information we have

$$nR = H(W) = I(W; Y^n) + H(W|Y^n). \tag{5.52}$$

Since the channel directly acts on X^n, we deduce that $p(y^n|x^n, w) = p(y^n|x^n)$, that is $W \to X^n \to Y^n$. Then, according to the properties of the Markov chains and in particular to DPI, from (5.52) it follows that

$$nR \leq I(X^n; Y^n) + H(W|Y^n). \tag{5.53}$$

By exploiting the lemmas proved above, from (5.53) we get

$$\begin{aligned} nR &\leq I(X^n; Y^n) + 1 + P_e^{(n)} \log(2^{nR} - 1) \\ &< nC + 1 + P_e^{(n)} nR. \end{aligned} \tag{5.54}$$

Dividing by n yields:

$$R < C + \frac{1}{n} + P_e^{(n)} R. \tag{5.55}$$

It follows that if $n \to \infty$ and $P_e^{(n)} \to 0$ then $R < C + \varepsilon$ for any arbitrarily small ε, i.e. $R \leq C$.
According to the direct channel coding theorem, n must tend to infinity so

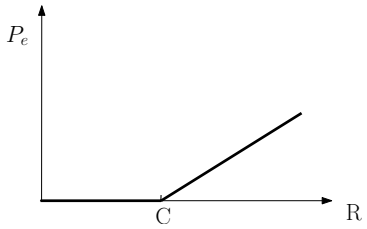

Figure 5.10: Asymptotic lower bound on P_e by varying R.

that $P_e^{(n)}$ can be made arbitrarily small. Therefore, if we want $P_e^{(n)} \to 0$ it's necessary that the rate R stays below capacity. This fact proves that $R < C$ is also a necessary condition for a rate R to be achievable.
From (5.55) there is another possible way through which we can show that if $R > C$ then $P_e^{(n)} \nrightarrow 0$. Let us rewrite (5.55) as follows

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}. \tag{5.56}$$

Joining this condition with the positivity of $P_e^{(n)}$ produces the asymptotical lower bound on P_e depicted in Figure 5.10. It's easy to see that if $R > C$ the probability of error is bounded away from 0 for large n. As a consequence, we cannot achieve an arbitrarily low probability of error at rates above capacity. □

5.2.4 Channel Coding in practice

The essence of the channel coding theorem is that, as long as $R < C$, it is possible to send information without affecting the reliability of the transmission. Hence, the noisiness of the channel does not limit the reliability of the transmission but only its rate. Moreover, Shannon proves that choosing the codes at random is asymptotically the best choice whatever the channel is. However, it is easy to deduce that for finite n the knowledge of the channel may help to choose a better code.
The problems we have to face with in practice are many. Hereinafter, we review the most common channels in order to compute the channel capacity C.

Evaluation of channel capacity

In order to evaluate the channel capacity of a given channel we have to solve the maximization

$$C = \max_{p(x)} I(X;Y), \tag{5.57}$$

for a given $p(y|x)$ and subject to the constraints on $p(x)$,

$$\begin{cases} p(x) \in [0,1] & \forall x \\ \sum_x p(x) = 1. \end{cases} \tag{5.58}$$

It's possible to prove that since $p(y|x)$ is fixed by the channel, the mutual information is a *concave function* of $p(x)$. Hence, a maximum for $I(X;Y)$ exists and is unique. However, being the objective function a nonlinear function, solving (5.57) is not easy and requires using methods of numerical optimization. There are only some simple channels, already introduced at the beginning of the chapter, for which it is possible to determine C analytically.

- *Noisy typewriter*

In this channel if we know the input symbol we have two possible outputs (the same or the subsequent symbol) with a probability 1/2 for each. Then, $H(Y|X) = 1$ and $\max I(X;Y) = \max(H(Y) - H(Y|X))) = \max(H(Y) - 1)$. The maximum of the entropy of the output source, which is $\log |\mathcal{Y}|$, can be achieved by using $p(x)$ distributed uniformly over all the inputs. Since the input and the output alphabet coincide, we have

$$C = \log |\mathcal{Y}| - 1 = \log |\mathcal{X}| - 1. \tag{5.59}$$

We deduce that, due to the action of the channel, we loose 1 information bit. Equivalently, the maximum rate of transmission is $C = \log \frac{|\mathcal{X}|}{2}$. This suggests that the intuitive idea of considering half symbols we proposed at the beginning of Section 5.2 is an optimum choice. It may come as a paradox that the value C is obtained by considering the inputs equally likely, but this is not necessarily the way according to which we have to take the inputs if we want to transmit at rate C. In fact, in this particular case, taking only non consecutive inputs permits to send information through the channel at the maximum rate C, without having to send n to infinity. This is not a contradiction; Shannon proposes a conceptually simple encoding and decoding scheme, this does not preclude the existence of better schemes, especially for finite n. What is certain is that the transmission rate cannot go beyond C.

• *BSC*

Even for this channel the maximization of the mutual information is straightforward, since we can easily compute the probability distribution $p(x)$ which maximizes $H(Y)$. As we already know from the analysis in Section 5.2.1, $C = \max(H(Y) - h(\varepsilon)) = 1 - h(\varepsilon)$, which is achieved when the input distribution is uniform.

• *BEC*

For the binary erasure channel (Figure 5.7) the evaluation of the capacity is a little bit more complex. Since $H(Y|X)$ is a characteristic of the channel and does not depend on the probability of the input, we can write

$$\begin{aligned} C &= \max_{p(x)}(H(Y) - H(Y|X)) \\ &= \max_{p(x)} H(Y) - h(\alpha). \end{aligned} \tag{5.60}$$

For a generic value of α the absolute maximum value for $H(Y)$ ($\log|\mathcal{Y}| = \log 3$) cannot be achieved for any choice of the input distribution. Then, we have to explicitly solve the maximization problem. Let $p_X(0) = \pi$ and $p_X(1) = 1 - \pi$. There are two ways for the evaluation of π. According to the first method, from the output distribution given by the triplet $p_Y(y) = (\pi(1-\alpha), \alpha, (1-\pi)(1-\alpha))$ we calculate the entropy $H(Y)$ and later maximize on π. The other method exploits the grouping property, yielding

$$\begin{aligned} H(Y) =& H_3(\pi(1-\alpha), \alpha, (1-\pi)(1-\alpha)) \\ =& H_2(\alpha, (1-\alpha)) + (1-\alpha)H_2(\pi, 1-\pi) = h(\alpha) + (1-\alpha)h(\pi). \end{aligned} \tag{5.61}$$

The maximum of the above expression is obtained when $h(\pi) = 1$, and then for $\pi = 1/2$. It follows that $C = h(\alpha) + (1-\alpha) - h(\alpha) = 1 - \alpha$. The result is expected since the BEC channel is nothing else that a noiseless binary channel which breaks down with a probability α; then, C can be obtained substracting to 1 the fraction of time the channel remains inoperative.

Construction of the codes

The channel coding theorem promises the existence of block codes that allow to transmit information at rates below capacity with arbitrarily small probability of error if the block length is large enough. The greatest problem

of channel coding is to find codes which allows in practice to transmit at rate close to C. Ever since the appearance of Shannon's paper, people have searched for such codes. In addition, usable codes should be "simple", so that they could be encoded and decoded easily. If we generated the codewords at random, according to Shannon's scheme, we would have to list all the codewords and send them to the receiver, requiring a huge amount of memory. Furthermore, we need a way to associate the messages we have to transmit and the codewords. Besides, since the code must be invertible, the codewords have to be distinct among themselves. Shannon overcomes this problem considering an asymptotical situation. Sending n to infinity is also what makes possible to use the jointly typical decoding as decoding rule at the receiver side. Such decoding scheme requires the receiver to check all the sequences which may have been sent in order to make the decision on the transmitted codeword. However, even if we consider a minimum distance algorithm it may require up to 2^{nR} evaluations.