# Introduction to Quantum Information Theory

Carlos Palazuelos
Instituto de Ciencias Matemáticas (ICMAT)
*carlospalazuelos@icmat.es*
Madrid, Spain

March 2013

# Contents

CHAPTER 1

# A comment on these notes

These notes were elaborated during the first semester of 2013, while I was preparing a course on quantum information theory as a subject for the PhD programme: *Investigación Matemática* at Universidad Complutense de Madrid. The aim of this work is to present an accessible introduction to some topics in the field of quantum information theory for those people who do not have any background on the field. With that goal in mind, these notes have been written so that no previous knowledge about quantum mechanics nor information theory is required to follow them. Actually, very basic concepts from probability theory, linear algebra and real analysis will be needed. Since this PhD course was expected to last no more than 20 hours, I saw myself forced to select only a few topics among many different possibilities. As a consequence of this, the presented course is divided into thee different parts.

The first part of the work, which is developed along Chapters 2, 3 and 4, is devoted to the introduction of the *postulates of quantum mechanics*, as well as to the presentation of some definitions and results which will be used along the rest of the notes. We will also explain the important protocols of *quantum teleportation* and *superdense coding* in Chapter 3. For this first part of the course, the basic reference that I faithfully followed is [**8**].

The second part of the course, which can be found in Chapter 5, is devoted to the introduction of *Bell inequalities* or, equivalently, *quantum nonlocality.* The aim of this chapter is to present the problem of quantum nonlocality in its simplest form and also to show how certain mathematical results in the context of functional analysis can be applied in the quantum scenario. In particular, it is explained how *the fundamental theorem in the metric theory of tensor products* developed by Grothendieck, fits in the context of quantum nonlocality.

Finally, the last part of the work, composed by Chapter 6 and Chapter 7, is devoted to giving an introduction to certain problems in classical information theory and to studying how these problems can be treated in the context of quantum mechanics. To this end, in Chapter 6 one can find the statements and proofs of the *noiseless* and *noisy Shannon's channel coding theorems* respectively, while Chapter 7 deals with these theorems in the quantum context. My basic reference for these last two chapters has been [**11**], from which most of the intuitive ideas have been drawn. Reference [**15**] has also been used.

CHAPTER 2

# Postulates of quantum mechanics

## 1. Postulate I and Postulate II

Postulate I. *Associated to any isolated physical system there is a complex Hilbert space $H$, known as the* state space *of the system. The system is completely described by its* state vector*, which is a unit vector in the state space.*

The state space of the system $H$ will depend on the specific physical system that we are studying. In this course we will restrict to finite dimensional Hilbert spaces. However, the fact that Postulate I is stated in this abstract way will allow us to develop an elegant general theory which does not depend on the specific physical system we are considering. Through the whole work, we will denote the $n$-dimensional complex Hilbert space by $\mathbb{C}^n$. According to Postulate I, in order to describe the state of a physical system of dimension $n$ we will simply give a unit vector $|\psi\rangle \in \mathbb{C}^n$. Note that we are using the ket-notation $|\psi\rangle$ to denote a particular element of the corresponding Hilbert space, whereas we will use bra-notation $\langle\psi|$ to consider the corresponding dual vector. In this way, the action of a bra $\langle\varphi|$ on a ket $|\psi\rangle$ is expressed by the standard inner product $\langle\varphi|\psi\rangle$. Moreover, this notation is also very convenient to express rank-one operators acting on a Hilbert space $\mathbb{C}^n$. Indeed, $|\psi\rangle\langle\varphi| : \mathbb{C}^n \to \mathbb{C}^n$ is defined as

$$|\psi\rangle\langle\varphi|(|\xi\rangle) = \langle\varphi|\xi\rangle|\psi\rangle \quad \text{for every} \quad |\xi\rangle \in \mathbb{C}^n.$$

The simplest quantum mechanical system is the *qubit*. In fact the qubit will play the same role in quantum information theory as the *bit* in classical information theory; it is the basic unit of information. Formally, it is the system whose associated vector space is a two dimensional Hilbert space. We will use notation $\{|0\rangle, |1\rangle\}$ for the canonical basis of $\mathbb{C}^2$. This basis is usually called *computational basis*. Then, while a classical bit can be in the state 0 or in the state 1, an arbitrary state for a qubit is a vector

$$|\varphi\rangle = a|0\rangle + b|1\rangle \quad \text{with} \quad a, b \in \mathbb{C}, \ |a|^2 + |b|^2 = 1.$$

We will often think of a qubit as a system that can be in the situations $|0\rangle$ or $|1\rangle$ and, in the case when $a \neq 0 \neq b$, we will say that the state is in a *superposition* of both situations. Note that there is a crucial difference between the possible states of a bit (which are just two: 0 or 1) and the possible states of a qubit (which are infinite!). This new situation will allow us to perform new protocols for quantum information processing. It is important to note that, although a given qubit can be in any superposition state $a|0\rangle + b|1\rangle$, whenever we "measure" the state of such a qubit, we will find it either in the state $|0\rangle$ or in the state $|1\rangle$ - as in the classical case - (with certain probabilities). However, although we cannot "see" the superposition phenomenon by measuring on the state, we will be able to use it! We will explain this point in more detail in Section 2. Apart from the states $|0\rangle$ and $|1\rangle$, the following two states will be very useful in the following.

$$|+\rangle = \frac{1}{\sqrt{2}}\big(|0\rangle + |1\rangle\big), \quad |-\rangle = \frac{1}{\sqrt{2}}\big(|0\rangle - |1\rangle\big).$$

Note that they also form an orthonormal basis of $\mathbb{C}^2$.

Postulate II. *The evolution of an isolated physical system (with associated Hilbert space $H$) is described by a* unitary transformation. *That is, if the state of the system at time $t_1$ is* *described by* $|\varphi_1\rangle$ *and the state of the system at time* $t_2 > t_1$ *is described by* $|\varphi_2\rangle$*, then there exists a unitary operator* $U \in B(H)$*, which depends only on the times* $t_1$ *and* $t_2$*, such that*

$$|\varphi_2\rangle = U|\varphi_1\rangle.$$

*Here, we denote by* $B(H)$ *the space of bounded operators on* $H$*.*

In fact, "a more physical statement" for the Postulate II should make use of the Hamiltonian operator defining the system.

Postulate II (twice). *The time evolution of the state of a closed quantum system is described by the Schrödinger equation,*

$$i\hbar\frac{d|\psi_t\rangle}{dt} = H|\psi_t\rangle.$$

*In this equation,* $\hbar$ *is the Planck's constant and* $H$ *is a fixed Hermitian operator known as the* Hamiltonian *of the closed system.*

Because the Hamiltonian is a hermitian operator it has a spectral decomposition

$$H = \sum_E E|E\rangle\langle E|,$$

with eigenvalues $E$ and corresponding normalized eigenvectors $|E\rangle$. The states $|E\rangle$ are usually called *energy eigenstates* or *stationary states*, and $E$ is the energy of the state $|E\rangle$. The lowest energy is known as *ground state energy* for the system, and the corresponding eigenstate is known as the *ground state.* The states $|E\rangle$ are called stationary states because their only change in time is of the form

$$|E\rangle \to \exp(-iEt/\hbar)|E\rangle.$$

What is the connection between Postulate II and Postulate II (twice)? The answer follows from the solution to Schrödinger's equation, which can be easily verified to be:

$$|\psi(t_2)\rangle = \exp\Big[\frac{-iH(t_2-t_1)}{\hbar}\Big]|\psi(t_1)\rangle = U(t_1,t_2)|\psi(t_1)\rangle,$$

where we define

$$U(t_1,t_2) := \exp\Big[\frac{-iH(t_2-t_1)}{\hbar}\Big].$$

It is not difficult to see that this operator is unitary and, furthermore, that any unitary operator $U$ can be realized in the form $U = \exp(iK)$ for some Hermitian operator $K$.

In these notes, we will adopt the point of view of Postulate II and we will not need to refer neither the Hamiltonian of the system nor the Schrödinger equation. Remarkably, any unitary $U \in B(\mathbb{C}^n)$ defines a certain evolution on the system with associated Hilbert space $\mathbb{C}^n$.

In the particular case of qubits, there are distinguished unitaries called $\sigma_0, \sigma_x, \sigma_y, \sigma_z$, or just $\mathbb{1}, X, Y, Z$. They are defined by

$$\sigma_0 = \mathbb{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad \sigma_1 = \sigma_x = X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix};$$

$$\sigma_2 = \sigma_y = Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}; \quad \sigma_3 = \sigma_z = Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

These matrices are called *the Pauli matrices.*

From a quantum computational point of view we can think of unitary matrices as quantum logical gates. Indeed, a classical circuit is a system formed by *wires* (to transmit the bits) and logic gates (acting on bits). In the same way we can think of a quantum circuit as a system formed by wires (to transmit qubits) and quantum logic gates (acting on qubits). This quantum gates are precisely the unitary matrices. In the case of one bit, we only have a non trivial gate which is the *NOT* one:

$$\begin{cases} 0 \to 1 \\ 1 \to 0. \end{cases}$$

However, we have seen before that Pauli matrices are examples of different and non trivial quantum gates. In fact, note that the Pauli matrix $X$ can be seen as the analogous quantum gate to the NOT in the classical context. Indeed, $X$ transforms a qubit in the state $a|0\rangle + b|1\rangle$ into the state $b|0\rangle + a|1\rangle$. In particular, it will send the state $|0\rangle$ to $|1\rangle$ and the state $|1\rangle$ to $|0\rangle$. Another interesting quantum gate is the *Hadamard gate*:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

## 2. Postulate III

Postulate III. *In a given physical system with associated Hilbert space $H$,* quantum measurements *are described by a collection $\{M_n\}_n \subset B(H)$ of* measurement operators*. The index $n$ refers to the measurement outcomes that may occur in the experiment. If the state of the quantum system is $|\psi\rangle$ immediately before the measurement, then the probability that result $n$ occurs is given by*

$$p(n) = \langle\psi|M_n^\dagger M_n|\psi\rangle,$$

*and the state of the system after the measurement is*

$$\frac{M_n|\psi\rangle}{\sqrt{\langle\psi|M_n^\dagger M_n|\psi\rangle}}.$$

*Measurement operators satisfy*

$$\sum_n M_n^\dagger M_n = I,$$

*needed for the probabilities to sum one.*

One of the simplest examples is the measurement of a qubit in the computational basis. This is defined by the measurement operators

$$M_0 = |0\rangle\langle 0| \quad \text{and} \quad M_1 = |1\rangle\langle 1|.$$

It is easy to see that both operators are selfadjoint (actually they are projections), $M_i^\dagger M_i = M_i^2 = M_i$ $(i = 1, 2)$ and $M_0 + M_1 = \mathbb{1}$. It is also easy to check that, when we measure the state $|\varphi\rangle = a_0|0\rangle + a_1|1\rangle$, the probability of obtaining the outcome $i$ is $|a_i|^2$ and the state after measurement in that case is $\frac{a}{|a|}|i\rangle$. We will see in the following that this state is "equivalent" to the state $|i\rangle$.

Indeed, if we consider the states $|\varphi\rangle$ and $e^{\imath\theta}|\varphi\rangle$ and we assume that we measure both states with a measurement $\{M_n\}_n$, then the probability of outcome $n$ is, in the second case,

$$\langle\varphi e^{-\imath\theta}|M_n^\dagger M_n|e^{\imath\theta}\varphi\rangle = \langle\varphi|M_n^\dagger M_n|\varphi\rangle.$$

Therefore both states are operationally identical.

**2.1. Distinguishability.** One of the typical problems in quantum information will be to distinguish two (or more) quantum states from each other. That is, we have a particle in one of several possible states and we want to find out in which of them the particle actually is. We will study this problem now in the simplest case: distinguish between two possible states.

Let us first assume that the states we want to distinguish, $|\varphi_1\rangle$ and $|\varphi_2\rangle$ are orthogonal. Then we can choose the measurement operators $M_i = |\varphi_i\rangle\langle\varphi_i|$ $(i = 1, 2)$ and $M_0 = \mathbb{1} - \sum_i M_i$. Note that all these operators are projections and trivially sum up to $\mathbb{1}$. Then, if $|\varphi\rangle$ is prepared in the state $|\varphi_i\rangle$ then

$$p(i) = \langle\varphi|M_i|\varphi\rangle = 1, \quad \text{and} \quad p(j) = 0, \quad \text{for every} \quad j \neq i.$$

Therefore, both states can be unambiguously distinguished.

Suppose now that we want to distinguish two states $|\varphi_1\rangle$ and $|\varphi_2\rangle$ which are not orthogonal. Let us prove that there is no way we can do that:

Assume there is a measurement $\{M_n\}_{n\in I}$ capable of distinguishing both states. In that case, we must be able to decompose $I = I_1 \cup I_2$ disjointly so that we can decide that the state is $|\varphi_i\rangle$ if the result of the measurement is $n_0 \in I_i$. Consider then the operators $E_i = \sum_{n\in I_i} M_n^\dagger M_n$.

We must have

$$\langle\varphi_i|E_i|\varphi_i\rangle = 1(i = 1, 2).$$

Since $E_1 + E_2 = \mathbb{1}$, we get $\langle\varphi_1|E_2|\varphi_1\rangle = 0$. Since $E_2$ is positive, we can write

$$0 = \langle\varphi_1|E_2|\varphi_1\rangle = \langle\varphi_1|\sqrt{E_2}\sqrt{E_2}|\varphi_1\rangle,$$

hence $\sqrt{E_2}|\varphi_1\rangle = 0$.

Since $|\varphi_1\rangle$ and $|\varphi_2\rangle$ are not orthogonal, we know that there exist $\alpha \neq 0 \neq \beta$ and $|\psi\rangle$ orthogonal to $|\varphi_1\rangle$ such that $|\alpha|^2 + |\beta|^2 = 1$ and

$$|\varphi_2\rangle = \alpha|\varphi_1\rangle + \beta|\psi\rangle.$$

Then we must have

$$\sqrt{E_2}|\varphi_2\rangle = \alpha\sqrt{E_2}|\varphi_1\rangle + \beta\sqrt{E_2}|\psi\rangle = \beta\sqrt{E_2}|\psi\rangle,$$

but this is a contradiction since

$$\|\beta\sqrt{E_2}|\psi\rangle\| \leq |\beta| < 1$$

and

$$\|\sqrt{E_2}|\varphi_2\rangle\| = 1.$$

**2.2. POVM Measurements.** In many cases, we will not be interested in the post-measurement state of our particle, but only in the probabilities of the different possible measurement outcomes. This leads us to the formalism of the so called *Positive Operator Valued Measurements* (POVM's). Suppose we have a measurement $\{M_n\}_n$ defined as in Postulate III. Then we can define the *positive* operators $E_n = M_n^\dagger M_n$. We have that $\sum_n E_n = \mathbb{1}$ and that the probability of obtaining outcome $m$ is

$$p(m) = \langle\varphi|E_m|\varphi\rangle.$$

Conversely, whenever we have a collection of positive operators $\{E_n\}_n$ such that $\sum_n E_n = \mathbb{1}$ we can define the measurement $\{M_n\}_n$, where $M_n = \sqrt{E_n}$.

**2.3. Projective Measurements.** In many applications, we will be very interested in a special case of measurements called *projective measurements.* These are measurements $\{M_n\}_n$ as in Postulate III with the additional property that each the $M_n$'s are orthogonal projections; that is, they are selfadjoint and verify

$$M_n M_m = \delta_{mn} M_n.$$

In this case, we can define an *observable* $M$ as the Hermitian operator

$$M = \sum_n n M_n.$$

With this notation, the average value of the measurement (on a state $|\varphi\rangle$) is

$$\sum_n np(n) = \sum_n n\langle\varphi|M_n^\dagger M_n|\varphi\rangle = \sum_n n\langle\varphi|M_n|\varphi\rangle = \langle\varphi|M|\varphi\rangle.$$

Conversely, if we consider a Hermitian operator $M$, we can consider its spectral decomposition and write it like

$$M = \sum_n \lambda_n P_n,$$

where each $P_n$ is a projection onto an eigenspace.

## 3. Postulate IV

Postulate IV. *The state space of a composite physical system is the tensor product of the state spaces of the component physical systems. Moreover, if system number $i$ is prepared in the state $|\varphi_i\rangle$ then the composite system is in the state $|\varphi_1\rangle \otimes \cdots \otimes |\varphi_n\rangle$.*

Most of the times, we will omit the tensor notation and we will just write $|\varphi_1\rangle|\varphi_2\rangle\cdots|\varphi_n\rangle$ to denote $|\varphi_1\rangle \otimes |\varphi_2\rangle \otimes \cdots \otimes |\varphi_n\rangle$.

An elementary but crucial observation to Postulate IV is that the tensor product of several Hilbert spaces $H_1 \otimes \cdots \otimes H_n$ contains elements $|\psi\rangle$ which are not elementary tensor products; that is, which cannot be written as $|\varphi_1\rangle\cdots|\varphi_n\rangle$ with $|\varphi_i\rangle \in H_i$. The elementary tensors correspond to those states which have been prepared independently by each party. On the other hand, a standard example of a non elementary two qubits state is the *EPR pair* or *Bell state* $|\psi\rangle \in \mathbb{C}^2 \otimes \mathbb{C}^2$, defined as

$$|\Phi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}.$$

Actually, this richer structure of the tensor product allows us to "define" the *quantum entanglement*, a behavior that seems to be at the root of many of the most surprising phenomena in quantum mechanics. Given a state $|\psi\rangle \in H_1 \otimes \cdots \otimes H_n$, we say that $|\psi\rangle$ is *entangled* if it cannot be written as an elementary tensor product. Note that, in particular, we need to have more than one system to talk about entangled states.

The canonical basis of $\mathbb{C}^2 \otimes \mathbb{C}^2$ is given by

$$\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}.$$

Another interesting basis for the two qubit states, the Bell basis, is given by

$$\begin{cases} |\Phi^+\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}, \\ |\Phi^-\rangle = \frac{|00\rangle-|11\rangle}{\sqrt{2}}, \\ |\Psi^+\rangle = \frac{|01\rangle+|10\rangle}{\sqrt{2}}, \\ |\Psi^-\rangle = \frac{|01\rangle-|10\rangle}{\sqrt{2}}. \end{cases}$$

Considering multiple qubits is not only important because of the appearance of the entanglement, but it also implies many more possibilities in difference senses. In particular, we have many new quantum gates (or evolutions) that we can perform. Note that, in the same way as for the elements in $H_1 \otimes \cdots \otimes H_n$, the unitaries acting on the tensor product of Hilbert spaces $U : H_1 \otimes \cdots \otimes H_n \to H_1 \otimes \cdots \otimes H_n$ can be very different from those defined as $U = U_1 \otimes \cdots \otimes U_n$ with $U_i : H_i \to H_i$ a unitary; which correspond to evolutions of the system given by independent evolutions in each subsystem.

We already pointed out that in the case of classical bits, the flip is the only non trivial gate. However, when we consider multiple bits we have some new interesting gates. Let us consider the case of two bits:

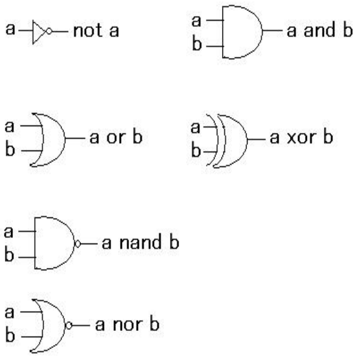

FIGURE 1. Some classical gates for two bits.

An important theoretical result is that any function on bits can be computed from the composition of NAND gates alone, which is thus known as a *universal gate*. By contrast, the XOR alone or even with NOT is not universal (just look at the parity!).

The prototypical multi(bi)-qubit quantum logic gate is the *controlled*-NOT or CNOT gate:

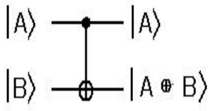

FIGURE 2. Controlled-NOT

$$U_{CN} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

If the control qubit $|A\rangle$ is 0, then the target qubit is not modified. If the control qubit is 1, then the target qubit is flipped. Another way to understand the CNOT gate is as a generalization of the XOR gate. Note, however, that there are some classical gates, like NAND or XOR, which cannot be understood as unitary gates in a sense similar to the way the quantum NOT gate represents the classical NOT gate. The reason is because these two gates are essentially irreversible. For example, given an output $A \oplus B$ of a XOR gate, it is not possible to determine what the inputs $A$ and $B$ were; there is a loss of information associated with the irreversible action of the XOR gate. On the other hand, being quantum gates described by unitary matrices, they can always be inverted by another quantum gate. Of course, there are many interesting quantum gates other than the controlled-NOT. However, this gate has the following remarkable universality property: Any multiple qubit logic gate may be composed from CNOT and single qubit gates.

Regarding measurements on a composite system we can consider those which are given by independent measurement in each party as a particular case of general measurements on the composite system. Let us consider for simplicity a setting described by two $n$-dimensional systems. Therefore, the system will be described by a quantum bipartite state $|\psi\rangle \in \mathbb{C}^n \otimes \mathbb{C}^n$. Let us assume that we perform a measurement with $K$ possible outputs on the first system and another measurement with $J$ possible outputs on the second system. Let us assume that those measurements are given by $\{M_a\}_{a=1}^K$ and $\{M'_b\}_{b=1}^J$ respectively. In this case, we will have a measurement on the composite system with $KJ$ possible outputs defined by $\{M_a \otimes M'_b\}_{a,b=1}^{K,J}$. Of course, this is nothing else than a particular, although very important, example of a general measurement on the composite system which will be given by $\{M_i\}_{i=1}^N$, where each $M_i$ is an operator acting on $\mathbb{C}^n \otimes \mathbb{C}^n$.

We finish this chapter by showing that projective measurements are as universal as general measurements, for as long as we allow for the use of *ancilla systems*. Suppose we have a physical system with state space $H$, and we want to perform a measurement $\{M_n\}_{n\in I}$ in it. To do this only with projective measurements, we introduce an auxiliary system (ancilla system) with state space $K$, where $K$ is a $|I|$-dimensional system with orthogonal basis $(|n\rangle)_{n\in I}$. Let $|0\rangle$ be a fixed state of $K$. Let

$$U : H \otimes [|0\rangle] \longrightarrow H \otimes K$$

be defined by

$$U|\varphi\rangle|0\rangle = \sum_{n\in I} M_n|\varphi\rangle|n\rangle.$$

Let us see that $U$ preserves inner products on $H \otimes [|0\rangle]$. Take $|\varphi\rangle$, $|\psi\rangle \in H$. Then, using the orthonormality of $(|i\rangle)_i$ and condition $\sum_i M_i^\dagger M_i = \mathbb{1}$, we have

$$\langle\varphi|\langle 0|U^\dagger U|\psi\rangle|0\rangle = \sum_{i,j}\langle\varphi|\langle i|M_i^\dagger M_j|\psi\rangle|j\rangle = \sum_i\langle\varphi|\langle i|M_i^\dagger M_i|\psi\rangle|i\rangle =$$

$$= \sum_i\langle\varphi|M_i^\dagger M_i|\psi\rangle = \langle\varphi|\psi\rangle.$$

It is an easy exercise now to see that in that case $U$ can be extended to an unitary operator (which we also call $U$)

$$U : H \otimes K \longrightarrow H \otimes K.$$

Now, we consider the projective measurement in the composite system $H \otimes K$ given by the projections $P_n = \mathbb{1}_H \otimes |n\rangle\langle n|$. The state we consider for the composite system is $U|\varphi\rangle|0\rangle = \sum_n M_n|\varphi\rangle|n\rangle$.

In that case, the probability of outcome $n$ taking place is

$$p(n) = \langle\varphi|\langle 0|U^\dagger P_n U|\varphi\rangle|0\rangle = \langle\varphi|M_n^\dagger M_n|\varphi\rangle.$$

This result is exactly as if we would have considered the measurement $\{M_n\}_n$ in the system $H$ acting on the state $|\varphi\rangle$. Note that the post-measurement state in the new case is

$$\frac{P_n U|\varphi\rangle|0\rangle}{\sqrt{\langle\varphi|\langle 0|U^\dagger P_n U|\varphi\rangle|0\rangle}} = \frac{M_n|\varphi\rangle|n\rangle}{\sqrt{\langle\varphi|M_n^\dagger M_n\varphi\rangle}}.$$

CHAPTER 3

# Some basic results

## 1. No-cloning theorem

The main motivation for this section is the following question: *Can we clone a classical/quantun unknown bit?*

There is a very easy way to see that the answer in the classical case, so in the case of bits, is affirmative. Indeed, this is a trivial application of the classical CNOT gate.

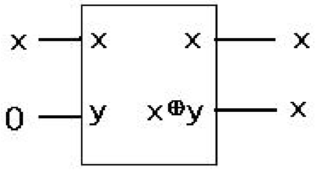

Figure 1. Classical circuit to copy an unknown bit.

If we start with the bit $x$ we want to clone (and we assume unknown) as a control bit and we start with the 0 target bit, we immediately obtain an input given by two copies of $x$. One could be tempted to think that a similar argument can be followed by using the quantum CNOT gate. Surprisingly, a similar argument does not work in the quantum context. In fact, the following result shows that it is not possible to clone an unknown quantum state.

Theorem 1.1. *It is not possible to perfectly clone an unknown quantum state using a unitary evolution.*

Proof. Suppose that we have a quantum machine with two inputs labeled by $A$ and $B$, as we had in the classical case. Input $A$, the *data input*, starts out in an unknown pure quantum state, $|\psi\rangle$. This is the state which must be be copied into input $B$, the *target input.* Let us assume that the target inputs is initiated in a pure state $|\phi\rangle$. Therefore, the initial state of the copy machine is

$$|\psi\rangle \otimes |\phi\rangle.$$

Some unitary evolution should act now on the system to transform it into

$$U(|\psi\rangle \otimes |\phi\rangle) = |\psi\rangle \otimes |\psi\rangle.$$

Let us suppose that this procedure works for two particular pure states $|\psi\rangle$ and $|\varphi\rangle$. Then, we will have

$$\begin{cases} U(|\psi\rangle \otimes |\phi\rangle) = |\psi\rangle \otimes |\psi\rangle, \\ U(|\varphi\rangle \otimes |\phi\rangle) = |\varphi\rangle \otimes |\varphi\rangle. \end{cases}$$

However, by using the fact that $U$ respects inner products we must have

$$\langle \psi, \varphi \rangle = |\langle \psi, \varphi \rangle|^2.$$

we deduce from here that $|\psi\rangle$ and $|\varphi\rangle$ are either the same state or orthogonal states. Thus cloning devices can only clone states which are orthogonal to one another and, therefore, a general quantum cloning device is impossible. □

What about the existence of a cloning device which is not given by a unitary evolution? It can be proved that even if one allows non-unitary cloning devices the perfect cloning of non-orthogonal pure states remains impossible. Most of the work during the last years have been devoted to study the existence of a cloning procedure (even for mixed states) in which we tolerate a certain lost of fidelity in the copied states. However, that is beyond the scope of this course.

## 2. Quantum teleportation

Quantum teleportation is one of the most representative communication protocols of quantum information theory. Let us assume that two people, say Alice and Bob, which are space separately share an EPR pair. Quantum teleportation is a communication protocol which allows Alice to send a qubit of information to Bob by just sending two classical bits. This is usually expressed by writing:

$$1 \text{ EPR} + 2 \text{ bits} \ \geq \ 1 \text{ qubit}.$$

In fact we should remark that Alice does not need to know her qubit in order to send it to Bob. Let us see how they must proceed. Let us assume that Alice's qubit is

$$|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle,$$

and we can assume that Alice does not know the values of $\alpha$ and $\beta$.

*Quantum teleportation algorithm:*

0. Alice and Bob share an EPR state $|\psi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$, so that the state of the whole system is
$$|\varphi_0\rangle = |\varphi\rangle|\psi\rangle = \frac{1}{\sqrt{2}}\Big[\alpha|0\rangle_A\big(|00\rangle_{A'B'} + |11\rangle_{A'B'}\big) + \beta|1\rangle_A\big(|00\rangle_{A'B'} + |11\rangle_{A'B'}\big)\Big].$$
1. Alice applies a CNOT on her part of $|\varphi_0\rangle$ to obtain
$$|\varphi_1\rangle = U_{A,A'} \otimes \mathbb{1}_B\big(|\varphi_0\rangle\big) = \frac{1}{\sqrt{2}}\Big[\alpha|0\rangle\big(|00\rangle + |11\rangle\big) + \beta|1\rangle\big(|10\rangle + |01\rangle\big)\Big].$$
2. Alice applies a Hadamard operation on $|\varphi\rangle$:
$$\begin{aligned}|\varphi_2\rangle &= \frac{1}{\sqrt{2}}\Big[\alpha\Big(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\Big)\big(|00\rangle + |11\rangle\big) + \beta\Big(\frac{|0\rangle - |1\rangle}{\sqrt{2}}\Big)\big(|10\rangle + |01\rangle\big)\Big] \\ &= \frac{1}{2}\Big[|00\rangle_{A,A'}\big(\alpha|0\rangle_B + \beta|1\rangle_B\big) + |01\rangle_{A,A'}\big(\alpha|1\rangle_B + \beta|0\rangle_B\big) \\ &\quad + |10\rangle_{A,A'}\big(\alpha|0\rangle_B - \beta|1\rangle_B\big) + |11\rangle_{A,A'}\big(\alpha|1\rangle_B - \beta|0\rangle_B\big)\Big].\end{aligned}$$

   Note that this expression breaks down into four terms. The first term has Alice's qubits in the state $|00\rangle$ and Bob's one in the state $|\varphi\rangle = \alpha|0\rangle_B + \beta|1\rangle_B$. Thus, if Alice measures in the computational basis
$$\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$$

and she obtains the output corresponding to $|00\rangle$, Bob's system will be in the state $|\varphi\rangle$. Similarly, from the previous expression we can read off Bob's postmeasurement state, given the result of Alice's measurement:

$$\begin{cases} 00 \mapsto |\varphi_3(00)\rangle := \alpha|0\rangle + \beta|1\rangle \\ 01 \mapsto |\varphi_3(01)\rangle := \alpha|1\rangle + \beta|0\rangle \\ 10 \mapsto |\varphi_3(10)\rangle := \alpha|0\rangle - \beta|1\rangle \\ 11 \mapsto |\varphi_3(11)\rangle := \alpha|1\rangle - \beta|0\rangle. \end{cases}$$

Depending on Alice's measuremet outcome, Bob's qubit will end up in one of these possible states. Of course, to know which state it is in, Bob must be told the result of Alice's measurement. Once Bob's has learned the measurement outcome, Bob can apply an appropiate quantum (on his qubit) to recover the state $|\varphi\rangle$. For example, in the case where the measuremnt output is 00, Bob doesn't need to do anything. If the measurement output is 01, so his qubit is in the state $\alpha|1\rangle + \beta|0\rangle$, Bob can recover the state $|\varphi\rangle$ by applying the $X$ gate to his state. In the case 10, Bob can fix up his state by applying the $Z$ gate. Finally, in the case 11, Bob should apply first an $X$ and after that a $Z$ gate.

**2.1. Some remarks on quantum teleportation.** There are two important questions about quantum teleportation which should be clarified before continuing.

a) *Are not we showing that we can transmit quantum states instantaneously?*
This would be peculiar because the theory of relativity states that nothing can travel faster than light. However, quantum teleportation does not allow for faster than light communication, because to complete the protocol Alice must transmit her measurement result to Bob over a classical communication channel. This classical channel is limited, of course, by the speed of light. It can be seen that without this classical communication, teleportation does not convey any information at all.

b) *Are not creating a copy of the state $|\varphi\rangle$ being teleported obtaining a contradiction with the no-cloning theorem?*
Again, this is just an illusion since after the teleportation process only the target qubit is left in the state $|\varphi\rangle$ and the original data qubit ends up in one of the computational basis state $|0\rangle$ or $|1\rangle$, depending upon the measurement result on the first qubit.

## 3. Superdense coding

Superdense coding is a communication protocol which allows, by assuming that Alice and Bob share an EPR pair, to transmit 2 classical bits of information by just sending 1 qubit. This is usually expressed by writing:

$$1 \text{ EPR} + 1 \text{ qubit } \geq 2 \text{ bits.}$$

That is, by sending the single qubit in her possession to Bob, it turns out that Alice can communicate two bits of classical information to Bob.

*Superdense coding algorithm:*

0. Alice and Bob share an EPR state $|\psi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$.

1. If Alice wants to send:
$$\begin{cases} 00 & \text{She does nothing} \\ 01 & \text{She applies a } Z \text{ gate to her qubit} \\ 10 & \text{She applies a } X \text{ gate to her qubit} \\ 11 & \text{She applies a } iY \text{ gate to her qubit.} \end{cases}$$
It is easy to see that the resulting states are respectively:
$$\begin{cases} 00 : |\psi\rangle \mapsto \frac{|00\rangle+|11\rangle}{\sqrt{2}} \\ 01 : |\psi\rangle \mapsto \frac{|00\rangle-|11\rangle}{\sqrt{2}} \\ 10 : |\psi\rangle \mapsto \frac{|10\rangle+|01\rangle}{\sqrt{2}} \\ 11 : |\psi\rangle \mapsto \frac{|01\rangle-|10\rangle}{\sqrt{2}}. \end{cases}$$
After this, Alice sends her part of the state to Bob so that he is in possession on the whole state. Note that the previous states form the Bell basis of $\mathbb{C}^2 \otimes \mathbb{C}^2$ and can therefore be distinguished by an appropriate quantum measurement.
2. By measurement the whole state (now in possession of Bob) in the Bell basis, Bob can determine which of the four possible bit string Alice sent.

CHAPTER 4

# The density operators formalism

## 1. Postulates of quantum mechanics: The density operators formalism

**1.1. Trace.** Before we introduce the density operator formalism, let us recall the definition of trace, and some of its properties. As in the previous chapter, we assume initially that our Hilbert space $H$ is finite dimensional. For every operator $A : H \longrightarrow H$ represented by a matrix $M$, we can define its *trace* $tr(A)$ as the sum of the elements of the diagonal of $M$. It is easy to see that, then, the trace is linear and cyclic, in the sense that $tr(AB) = tr(BA)$. From this last property it follows that, for every unitary operator $U \in B(H)$,

$$tr(UAU^\dagger) = tr(U^\dagger UA) = tr(A).$$

Therefore, the trace of an operator is well defined and does not depend on the chosen basis.

The following property of the trace is very useful. Let $|\varphi\rangle \in H$ be a unit vector and $|\varphi\rangle\langle\varphi| : H \longrightarrow H$ be the projection on the direction on $|\varphi\rangle$. Let $A \in B(H)$ be an arbitrary operator. We want to evaluate $tr(A|\varphi\rangle\langle\varphi|)$. To do this, first we extend $|\varphi\rangle$ to a basis $(|i\rangle)$ of $H$, where $|\varphi\rangle = |1\rangle$. Then, we have

$$tr(A|\varphi\rangle\langle\varphi|) = \sum_i \langle i|A|\varphi\rangle\langle\varphi|i\rangle = \langle\varphi|A|\varphi\rangle.$$

**1.2. Density operators.** So far we have described the state of a physical system as a unit vector in the Hilbert space $H$. There is an equivalent description where states are no longer elements in the Hilbert space, but trace class operators on it. This last description offers advantages in certain problems, specially (but not only) when dealing with real experiments systems where noise is always present. We describe this in the following.

The situation we often face is that we will not know that "our system is in a state $|\varphi\rangle$, but rather we will know that our system is in one of the states $|\varphi_i\rangle$, with probability $p_i$ respectively". Therefore, we would like to consider something like the "state"

$$\sum_i p_i|\varphi_i\rangle,$$

where the $p_i$'s are positive numbers verifying $\sum_i p_i = 1$. The problem is that this is not a state anymore, since it is not unitary. A way to circumvent this difficulty is to associate each state $|\varphi\rangle$ to the operator (rank-one projection) $|\varphi\rangle\langle\varphi| \in S_1(H)$, where $S_1(H)$ is the Banach space of the operators $T : H \longrightarrow H$ with finite trace norm. So now, if we have a state in the previous situation, we can describe it with the positive trace one operator

$$\rho = \sum_i p_i|\varphi_i\rangle\langle\varphi_i|.$$

$\rho$ is called the *density operator* or *density matrix.*

To see that $\rho$ is indeed positive note that, for any $|\psi\rangle \in H$,

$$\langle\psi|\rho|\psi\rangle = \sum_i p_i\langle\psi|\varphi_i\rangle\langle\varphi_i|\psi\rangle = \sum_i p_i|\langle\varphi_i|\psi\rangle|^2 \geq 0.$$

On the other hand, the fact that $tr(\rho) = 1$ is trivial from the linearity of the trace and the fact that $tr(|\varphi_i\rangle\langle\varphi_i|) = 1$ for every $i$.

Conversely, assume we have a positive trace one operator $\rho \in S_1(H)$. Being positive, $\rho$ admits an spectral decomposition

$$\rho = \sum_j \lambda_j |j\rangle\langle j|,$$

where the eigenvectors $|j\rangle$ are orthogonal, with norm 1, and the eigenvalues $\lambda_j$ are real, positive and verify $\sum_j \lambda_j = 1$ (because $\rho$ has trace 1). Therefore, we can see the numbers $\lambda_j$ as the probabilities of our system being in the state $|j\rangle$.

Note also that if we now have our system with probability $q_j$ described by the density operator $\rho_j = \sum_i p_{ij}|\varphi_{ij}\rangle\langle\varphi_{ij}|$ then

$$\rho = \sum_j q_j \rho_j = \sum_{ij} q_j p_{ij} |\varphi_{ij}\rangle\langle\varphi_{ij}|$$

is again a density operator, and it describes our system. In particular, the set of density operators is a convex.

The postulates of quantum mechanics can be equivalently stated in terms of density operators.

Postulate I'. *Associated to any isolated physical system there is a complex Hilbert space known as the* state space *of the system. The system is completely described by its* density operator*, which is a positive operator $\rho \in S_1(H) \subset B(H)$ with trace one. If the system is in the state $\rho_i$ with probability $p_i$, then the density operator for the system is $\sum_i p_i \rho_i$.*

We will use the notation *pure states* for the states of the form $|\varphi\rangle\langle\varphi|$ and *mixed states* for states of the form $\rho = \sum_i p_i |\varphi_i\rangle\langle\varphi_i|$.

The evolution of a system $H$ is, like before, given by unitaries on $H$. To see how they act on $\rho = \sum_i p_i|\varphi_i\rangle\langle\varphi_i|$, just notice that if the system initially is in the state $|\varphi_i\rangle$ with probability $p_i$, then after the evolution given by $U$ it will be in state $U|\varphi_i\rangle$ with probability $p_i$, hence the associated density operator will be

$$\sum_i p_i |U\varphi_i\rangle\langle\varphi_i U^\dagger| = U\left(\sum_i p_i|\varphi_i\rangle\langle\varphi_i|\right)U^\dagger = U\rho U^\dagger.$$

Therefore, the second postulate now says

Postulate II'. *The evolution of an isolated physical system (with associated Hilbert space $H$) is described by a* unitary transformation*. That is, if the state of the system at time $t_1$ is described by $\rho_1$ and the state of the system at $t_2 > t_1$ is described by $\rho_2$, then there exist a unitary operator $U \in B(H)$, which depends only on the times $t_1$ and $t_2$, such that*

$$\rho_2 = U\rho_1 U^\dagger.$$

As for the measurements, suppose we measure with a measurement $\{M_n\}$ a mixed state $\rho = \sum_i p_i|\varphi_i\rangle\langle\varphi_i|$. If the initial state is $|\varphi_i\rangle$ then the probability of outcome $n$ taking place is

$$p(n|i) = \langle\varphi_i|M_n^\dagger M_n|\varphi_i\rangle = tr(M_n^\dagger M_n|\varphi_i\rangle\langle\varphi_i|).$$

Therefore, the total probability of outcome $n$ is

$$p(n) = \sum_i p(n|i)p_i = \sum_i p_i tr(M_n^\dagger M_n|\varphi_i\rangle\langle\varphi_i|) = tr(M_n^\dagger M_n\rho).$$

With similar reasonings, we can see that the post-measurement state of the system when outcome $n$ has taken place is

$$\frac{M_n\rho M_n^\dagger}{tr(M_n\rho M_n^\dagger)}.$$

That is, our third postulate in this formalism reads

POSTULATE III'. *In a given physical system with associated Hilbert space $H$,* quantum measurements *are described by a collection $\{M_n\}_n \subset B(H)$ of* measurement operators. *The index $n$ refers to the measurement outcomes that may occur in the experiment. If the state of the quantum system is $\rho$ immediately before the measurement, then the probability that result $n$ occurs is given by*

$$p(n) = tr(M_n^\dagger M_n \rho),$$

*and the state of the system after the measurement is*

$$\frac{M_n \rho M_n^\dagger}{tr(M_n \rho M_n^\dagger)}.$$

*Measurement operators satisfy*

$$\sum_m M_n^\dagger M_n = I,$$

*needed for the probabilities to sum one.*

A simple consequence of the separate linearity of tensor products is

POSTULATE IV'. *The state space of a composite physical system is the tensor product of the state spaces of the component physical systems. Moreover, if system number $i$ is prepared in the state $\rho_i$ then the composite system is in the state $\rho_1 \otimes \cdots \otimes \rho_n$*

## 2. Partial trace

Suppose we have a composite physical system made up of the subsystems $H_A$ and $H_B$. Then, the system has associated the Hilbert space $H_A \otimes H_B$ and the state of the system is described by the density operator $\rho_{AB}$. Sometimes we need to describe the "A side" of our state. For that, we define the partial trace $tr_B$ as the linear operator

$$tr_B : S_1(H_A \otimes H_B) \longrightarrow S_1(H_A)$$

defined on elementary tensors by

$$tr_B(|a_1\rangle\langle a_2| \otimes |b_1\rangle\langle b_2|) = |a_1\rangle\langle a_2| tr(|b_1\rangle\langle b_2|),$$

where $tr(|b_1\rangle\langle b_2|)$ is the usual trace in $H_B$, hence equal to $\langle b_2|b_1\rangle$.

Let us see some examples of the action on the partial trace. Suppose the simplest case, where $\rho_{AB} = \rho_A \otimes \rho_B$. Then $tr_B(\rho_{AB}) = \rho_A tr(\rho_B) = \rho_A$, which is the result we would expect.

To see that things are in general not so simple, consider the EPR state

$$|\varphi\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}.$$

Its associated density operator is

$$\rho_{AB} = \left(\frac{|00\rangle + |11\rangle}{\sqrt{2}}\right)\left(\frac{\langle 00| + \langle 11|}{\sqrt{2}}\right) = \frac{|00\rangle\langle 00| + |11\rangle\langle 00| + |00\rangle\langle 11| + |11\rangle\langle 11|}{2}.$$

Then

$$\begin{aligned}\rho_A = tr_B(\rho_{AB}) &= \frac{tr_B(|00\rangle\langle 00|) + tr_B(|11\rangle\langle 00|) + tr_B(|00\rangle\langle 11|) + tr_B(|11\rangle\langle 11|)}{2} \\ &= \frac{|0\rangle\langle 0|\langle 0|0\rangle + |1\rangle\langle 0|\langle 0|1\rangle + |0\rangle\langle 1|\langle 1|0\rangle + |1\rangle\langle 1|\langle 1|1\rangle}{2} = \frac{|0\rangle\langle 0| + |1\rangle\langle 1|}{2} = \frac{\mathbb{1}}{2}.\end{aligned}$$

Note that, in this case, the original state $\rho_{AB}$ is a pure state, about which we have maximal knowledge, whereas $\rho_A$ is a mixed state, and indeed the identity, which is equivalent to not knowing anything!!

Let us show that the partial trace gives the "correct" statistics. Suppose we have a system $A$ where we measure an observable $M = \sum_m m P_m \in B(H)$. If we consider now a composite system $A \otimes B$, then the corresponding observable in this system is $M \otimes \mathbb{1}_B = \sum_m mP_m \otimes \mathbb{1}_B$, in the sense that, for any $|\varphi\rangle|\psi\rangle \in A \otimes B$, the probability of obtaining outcome $m$ is

$$p(m) = \langle \varphi \otimes \psi | P_m \otimes \mathbb{1}_B | \varphi \otimes \psi \rangle = \langle \varphi | P_m | \varphi \rangle \langle \psi | \mathbb{1}_B | \psi \rangle = \langle \varphi | P_m | \varphi \rangle,$$

which coincides with the probability of obtaining outcome $m$ if we measure only in the system $A$.

Consider now again both systems $A$ and $B$, and an observable $M$ in the system $A$. Suppose we have the system in the state $\rho_{AB}$ and we want to find a state $\rho_A$ for system $A$ which verifies that the average value we obtain when measure $\rho_A$ with $M$ coincides with the average value we obtain when we measure $\rho_{AB}$ with $M \otimes \mathbb{1}_B$. It is very easy to see that

$$tr(M\rho_A) = tr_B((M \otimes \mathbb{1}_B)\rho_{AB})$$

and, moreover, it can be proved that the partial trace is the only linear function verifying this.

## 3. The Schmidt decomposition and purifications

We state the following theorem without proof, which can be found in ([**8**, Section 2.5]).

THEOREM 3.1 (Schmidt decomposition). *Suppose $|\psi\rangle$ is a pure state of a composite system $AB$. Then, there exist orthonormal states $|i_A\rangle$ for system $A$, and orthonormal states $|i_B\rangle$ for system $B$ such that*

$$|\psi\rangle = \sum_i \lambda_i |i_A\rangle|i_B\rangle,$$

*where $\lambda_i$ are non-negative real numbers satisfying $\sum_i \lambda_i^2 = 1$ known as* Schmidt coefficients.

This result is very useful. As a taste of its power, consider the following consequence: let $|\psi\rangle$ be a pure state of a composite system, $AB$. Then, by the Schmidt decomposition $\rho_A = \sum_i \lambda_i^2 |i_A\rangle\langle i_A|$ and $\rho_B = \sum_i \lambda_i^2 |i_B\rangle\langle i_B|$, so the eigenvalues of $\rho_A$ and $\rho_B$ are identical, namely $\lambda_i^2$ for both density operators. Many important properties of quantum systems are completely determined by the eigenvalues of the reduced density operator of the system, so for a pure state of a composite system such properties will be the same for both systems.

The basis $|i_A\rangle$ and $|i_B\rangle$ are called *Schmidt bases* for $A$ and $B$, respectively, and, the number of non-zero values $\lambda_i$ is called the *Schmidt number* for the state $|\psi\rangle$. The Schmidt number is an important property of a composite quantum system, which in some sense quantifies the amount of entanglement between systems $A$ and $B$. To get an idea about this, consider the following important property: the Schmidt number is preserved under unitary transformations on system $A$ or system $B$ alone (the proof is really easy).

As a consequence of the above theorem we can prove that given any state $\rho_A$ of a quantum system $A$, it is possible to introduce another system $R$, and to define a pure state $|AR\rangle$ for the joint system $AR$ such that $\rho_A = tr_R(|AR\rangle\langle AR|)$. This is a purely mathematical procedure, known as *purification*, which allows us to associate pure states with mixed states. To prove the purification, suppose the state $\rho_A$ has an orthonormal decomposition $\rho_A = \sum_i p_i |i_A\rangle\langle i_A|$. We introduce now a system $R$ which has a state space with orthonormal basis $\{|i_R\rangle\}_i$ and we define the pure state for the combined system

$$|AR\rangle = \sum_i \sqrt{p_i} |i_A\rangle|i_R\rangle.$$

It is now very easy to verify that, indeed, $tr_R(|AR\rangle\langle AR|) = \rho_A$.

## 4. Definition of quantum channel and its classical capacity

**4.1. Classical channels.** Let us consider a scenario formed by a sender (Alice), a receiver (Bob) and a noisy channel to send information from the first to the second. As one would expect, a classical channel is a map sending (string of) bits. However, in order to "describe" noisy channels, we must consider some "errors" occurring with certain probabilities. As an illustrative example, let us consider the *binary symmetric channel* $\mathcal{N}$ which acts on a single bit and it is defined as follows.

$$\mathcal{N} : \begin{cases} 0 \quad \rightarrow & \begin{cases} 0 \text{ with probability } 1-p \\ 1 \text{ with probability } p, \end{cases} \\ 1 \quad \rightarrow & \begin{cases} 0 \text{ with probability } p \\ 1 \text{ with probability } 1-p. \end{cases} \end{cases}$$

Note that this channel actually sends bits to probability distributions on bits. Therefore, it is natural to define a classical channel as a map preserving probability distributions. Formally, a *classical channel* is defined as a (point-wise) positive linear map $\mathcal{N} : \mathbb{R}^n_A \rightarrow \mathbb{R}^n_B$ verifying $\sum_{i=1}^n \big(\mathcal{N}(P)\big)_i = \sum_{i=1}^n p_i$ for any $P = (p_i)_{i=1}^n \in \mathbb{R}^n$. In fact, with this definition at hand we see that the symbols sent by the channel $\mathcal{N}$ are not relevant, but the probability distribution describing the channel is what really matters. We can then understand a channel mapping an alphabet of $n$ letters $\{a_1, \cdots, a_n\}$ into another alphabet of m letters $\{b_1, \cdots, b_m\}$. Then, the channel is completely determined by the matrix $\big(P(b_j|a_i)\big)_{i,j}$, which can be seen as a linear map $\mathcal{N} : \mathbb{R}^n_A \rightarrow \mathbb{R}^m_B$, where $P(b_j|a_i)$ is the probability of obtaining the output $b_j$ where we send the input $a_i$ through the channel for every $i, j$. Note that, in order this channel to be well defined, we need to impose that $\sum_j P(b_j|a_i) = 1$ for every $i$, which is equivalent to the property stated above, when we gave the formal definition of a channel. Finally, in order to emphasize the pres ervation of probability distributions by a channel, we will denote a channel as above by $\mathcal{N} : \ell_1^n \rightarrow \ell_1^m$ (where $\ell_1^k = (\mathbb{R}^k, \|\cdot\|_1)$). In this way, if we are dealing with a channel acting on $n$-bit strings, we will denote $\mathcal{N} : \ell_1^{2^n} \rightarrow \ell_1^{2^n}$.

Following Shannon's ideas (**[13]**) the *capacity of a channel* is defined as an asymptotic ratio:

$$\frac{\text{number of transmitted bits with an } \epsilon \rightarrow 0 \text{ error}}{\text{number of required uses of the channel in parallel}}.$$

More precisely, given a channel $\mathcal{N} : \ell_1^n \rightarrow \ell_1^n$, its capacity is defined as

$$C_c(\mathcal{N}) := \lim_{\epsilon \rightarrow 0} \limsup_{k \rightarrow \infty} \Big\{ \frac{m}{k} : \exists_{\mathcal{A}} \exists_{\mathcal{B}} \text{ such that } \|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}\| < \epsilon \Big\}.$$

Here, $\mathcal{A} : \ell_1^{2^m} \rightarrow \bigotimes^k \ell_1^n$ and $\mathcal{B} : \bigotimes^k \ell_1^n \rightarrow \ell_1^{2^m}$ are channels and $\mathcal{N}^{\otimes_k} : \bigotimes^k \ell_1^n \rightarrow \bigotimes^k \ell_1^n$ denotes the use of $k$ times the channel in parallel. The composition $\mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}$ represents the protocol in which Alice encodes a message, sends this information using $k$ times the channel in parallel to Bob and Bob decodes the information he receives. The following diagram shows the situation:

$$\begin{tikzcd}
\bigotimes^k \ell_1^n \arrow[r, "\mathcal{N}^{\otimes_k}"] & \bigotimes^k \ell_1^n \arrow[d, "\mathcal{B}"] \\
\ell_1^{2^m} \arrow[u, "\mathcal{A}"] \arrow[r, "\mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}"] & \ell_1^{2^m}
\end{tikzcd}$$

and we want to have $\|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}\| < \epsilon$.

**4.2. Quantum channels.** As we have explained in Chapter 2, the basic unit in quantum information theory is the qubit: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \in \mathbb{C}^2$. Therefore, it is natural that quantum channels will allow to send strings of qubits. However, as in the classical case, in order to deal with noisy channels we are actually interested in probability distributions on string of qubits. This leads us to consider density operators: $\rho = \sum_{i=1}^N p_i|\psi_i\rangle\langle\psi_i|$. Then, we define a *quantum channel* as a completely positive and trace preserving linear map $\mathcal{N} : M_n \to M_n$. We recall that the map $T : B(H) \to B(K)$ is said *completely positive* if the map

$$1_{M_N} \otimes T : M_N(B(H)) \to M_N(B(K))$$

is positive for every natural number $N$ [1]. Since a quantum channel $\mathcal{N}$ acts on density operators it is natural to denote a channel by $\mathcal{N} : S_1^n \to S_1^n$, where $S_1^n$ denotes the space of trace class operators acting on $\mathbb{C}^n$. In fact, if the channel is acting on $n$-qubits states we must write $\mathcal{N} : S_1^{2^n} \to S_1^{2^n}$. Note that this comment is just about notation since $S_1^n = M_n$ algebraically. However, this notation becomes important if the Hilbert space considered is infinite dimensional, since in this case $M_n$ should be replaced by $B(H)$ and $S_1^n$ by $S_1(H)$. In the infinite dimensional case, both spaces are known not to be equal. In the following, we will use both notations $\mathcal{N} : S_1^n \to S_1^n$ and $\mathcal{N} : S_1(H_A) \to S_1(H_B)$ for a quantum channel.

The following result characterizes the form of quantum channels.

Theorem 4.1. *Given two Hilbert spaces $H_A$ and $H_B$ and a linear map $\mathcal{N} : S_1(H_A) \to S_1(H_B)$, the following conditions are equivalent:*

1. *$\mathcal{N}$ is completely positive and trace preserving.*
2. *There exist a Hilbert space $H_D$ and an isometry $V : H_A \to H_B \otimes H_D$ ($V^*V = 1_{H_A}$), such that*

$$\mathcal{N}(\rho) = tr_D(V\rho V^*) \tag{4.1}$$

*for every state $\rho$ in $S_1(H_A)$. Here, $tr_D$ denotes the trace over the system $D$.*

3. *There exist operators $E_1, \cdots, E_D$ in $B(H_A, H_B)$ verifying $\sum_{i=1}^D E_i^* E_i = \mathbb{1}_{B(H_A)}$ and such that*

$$\mathcal{N}(\rho) = \sum_{i=1}^D E_i \rho E_i^* \tag{4.2}$$

*for every state $\rho$ in $S_1(H_A)$. The operators $E_i$'s are usually called Krauss operators.*

Proof. The implication 1. ⇒ 2. is a direct consequence of Steinspring dilation theorem and the fact that $\mathcal{N}$ is a completely positive and trace preserving map if and only if $\mathcal{N}^*$ is a completely positive and unital map. Then, Steinspring dilation theorem (see for instance [**9**, Theorem 4.1]) states that there exist a Hilbert space $H_D$ and an isometry $V : H_A \to H_B \otimes H_D$ ($V^*V = 1_{H_A}$) such that

$$\mathcal{N}^*(A) = V^*(A \otimes 1_{H_D})V$$

for every $A \in B(H_B)$. The statement 2. follows trivially by taking adjoints.

In order to show 2. ⇒ 3., let us write the isometry $V : H_A \to H_B \otimes H_D$ by

$$V = \sum_{i=1}^D E_i \otimes |i\rangle,$$

[1] The requirement of completely positivity in the definition of quantum channels is explained by the fact that our map must be a channel when we consider our system as a physical subsystem of an amplified one (with an environment) and we consider the map $1_{Env} \otimes \mathcal{N}$.

where $D$ is the dimension of $H_D$, $(|i\rangle)_{i=1}^D$ is an orthonormal basis of $H_D$ and $E_i \in B(H_A, H_B)$ for every $i$. The fact that $V^*V = 1_{H_A}$ implies that $\sum_{i=1}^D E_i^* E_i = \mathbb{1}_{B(H_A)}$. On the other hand, it is very easy to check that $\mathcal{N}(\rho) = tr_D(V\rho V^*) = \mathcal{N}(\rho) = \sum_{i=1}^D E_i \rho E_i^*$ for every state $\rho$.

Finally, in order to prove 3. ⇒ 1., one can check that a map defined by Equation (4.1) is completely positive and trace preserving. □

Equation (4.1) has a nice interpretation in terms of the evolution of non isolated systems. Indeed, note that we can trivially extend the isometry $V : H_A \to H_B \otimes H_D$ to a unitary operator $U : H_A \otimes H_D \to H_B \otimes H_D$ and replace Equation (4.1) by

$$\mathcal{N}(\rho) = tr_D\big(U(\rho \otimes |0\rangle\langle 0|)U^*\big)$$

for every $\rho$, where $|0\rangle\langle 0|$ is an arbitrary fixed state in $B(H_D)$. This formulation links with our Postulate II about the evolution of isolated physical systems. Although we have introduced a quantum channel as a map sending quantum information, so qubits, this is nothing else than a map which describes an "arbitrary evolution" of quantum systems. The previous theorem says that any evolution can be seen as an evolution of an isolated system if we consider the composite system formed by ours and the environment. However, if we want to focus on our particular system, we will need to trace out the environment once the evolution has happened.

On the other hand, Equation (4.2) has a very nice interpretation in terms of the third postulate of quantum mechanics. Indeed, the condition $\sum_{i=1}^D E_i^* E_i = \mathbb{1}$ is exactly the property to be verified by any measurement. In particular, we know that the elements $\rho_i = \frac{E_i \rho E_i^*}{tr(E_i \rho E_i^*)}$ are states and we see that we can understand the action of the channel as

$$\mathcal{N}(\rho) = \sum_{i=1}^D p_i \rho_i,$$

where $p_i = tr(E_i \rho E_i^*)$. That is, $\mathcal{N}$ sends the state $\rho$ to the states $\rho_i$ with probability $p_i$. This allows us to understand a quantum channel as a generalization of a classical channel.

It is important to understand that a quantum channel can be used to send both, classical and quantum information. In Chapter 6 and Chapter 7 we will study the classical capacity of classical channels and the classical capacity of quantum channels respectively. In fact, even in the study of the classical capacity of quantum channels one could consider the case where the sender and the receiver are allowed to use an entangled state in the information protocol. Then, we talk about *classical capacity with assisted entanglement*. However, this context as well as the study of the quantum capacity of quantum channels are beyond the scope of these notes. The classical capacity of a quantum channel is defined in a similar manner to the case of classical channels:

$$\lim_{\epsilon \to 0} \limsup_{k \to \infty} \left\{ \frac{m}{k} : \exists_{\mathcal{A}}, \exists_{\mathcal{B}} \text{ such that } \|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}\| < \epsilon \right\}.$$

Here, $\mathcal{A} : \ell_1^{2^m} \to \otimes^k S_1^n$ will be a channel representing Alice's encoding from classical information to quantum information. On the other hand, Bob will decode the information he receives from Alice via the $k$ times uses of the channel, $\mathcal{N}^{\otimes_k}$, by means of another channel $\mathcal{B} : \otimes^k S_1^n \to \ell_1^{2^m}$. The following diagram represents the situation:

$$\begin{tikzcd}
\bigotimes^k S_1^n \arrow[r, "\mathcal{N}^{\otimes_k}"] & \bigotimes^k S_1^n \arrow[d, "\mathcal{B}"] . \\
\ell_1^{2^m} \arrow[u, "\mathcal{A}"] \arrow[r, "\mathcal{B}\circ\mathcal{N}^{\otimes_k}\circ\mathcal{A}"] & \ell_1^{2^m}
\end{tikzcd}$$

We look for the condition: $\|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}\| < \epsilon$.

CHAPTER 5

# Quantum nonlocality

Since the birth of quantum mechanics many authors have questioned this theory. Though they accepted that it is a *useful* theory which allows us to predict the physic laws in a very precise way, some of the most important scientists of the 20th century showed some skepticism about this theory owing to its nondeterministic nature. In fact, there have existed some alternative theories which tried to avoid the assumption that *uncertainty is intrinsic in Nature*, as it is assumed by quantum mechanics. Maybe, the most relevant theories have been those based on *hidden variable models*. As we have explained in the first chapter, quantum mechanics assumes that, in order to describe a physical system, we can just deal with a certain object (vector state) which contains *all the information that we can obtain about the system.* The impossibility of obtaining a more accurate information about it does not depend on our precision, but it is intrinsic in Nature. Contrary to this assumption, the hidden variable models propose that such an ignorance about Nature is due to our own restrictions. These models assume that there exists a *hidden probability* over the "possible sates of the world" that we cannot know - this is the way we model classically our uncertainty -. However, once one of these states is fixed, we are in a completely deterministic situation. That is, our uncertainty about Nature can be understood as a classical average over deterministic states.

In 1935 A. Einstein, B. Podolsky and N. Rosen proposed an experiment ([**4**]) whose aim was to prove the non completeness of quantum mechanics as a model of Nature. However, it took almost 30 years to understand that the apparent dilemma presented in [**4**] could be formulated in terms of assumptions which naturally lead to a refutable prediction. Bell showed that the assumption of a local hidden variable model implies some inequalities on the set of correlations obtained in a certain measurement scenario and that these inequalities, since then called *Bell inequalities*, are violated by certain quantum correlations produced with an entangled state ([**2**]).

Though initially discovered in the context of foundations of quantum mechanics, violations of Bell inequalities, commonly known as *quantum nonlocality*, are nowadays a key point in a wide range of branches of quantum information science. In particular, nonlocal correlations provide the quantum advantage in communication complexity and information theoretical protocols as well as in the security of quantum cryptography protocols.

In the following section we will explain Bell's result (in a modern language). After that, we will explain how nonlocality can be understood from a functional analysis point of view. In particular, we will briefly explain how *The fundamental theorem in the metric theory of tensor products*, developed by Grothendieck, perfectly fits in this context.

## 1. Bell's result: Correlations in EPR

We forget for a moment about quantum mechanics and we perform the following mental experiment. Charlie prepares two particles, in whatever way he wants, and he sends one of these particles to Alice and the other to Bob. Upon receiving her particle, Alice measures either property $Q$ or property $R$ of the particle, and assume that these measurements can only take the two values $\pm 1$. Bob does the same with his particle, and let us call $S, T$ to the properties he measures, again with the possible outcomes $\pm 1$. We assume that Alice and Bob can perform their measurements in a casually disconnected manner. That is, sufficiently simultaneously and far apart that the outcome of Alice's measurement can not influence in Bob's measurement and viceversa[1]. Let us also assume that Charlie can prepare similar pair of particles once and again and we can repeat the experiment as many times as we want. Then, we can talk about the probability distributions defined by the possible results of the experiments.

Let us consider the number

$$QS + RS + RT - QT = (Q + R)S + (R - Q)T.$$

From a local and deterministic point of view, it is clear that either $(Q+R)$ or $(R-Q)$ is 0 and

$$QS + RS + RT - QT = \pm 2.$$

Let us first assume that Nature can be explained by a Local Hidden Variable Model. The locality hypothesis means exactly what we have just explained about the possibility of Alice and Bob to perform their measurements in a casually disconnected manner. In particular, the special relativity theory implies this hypothesis. On the other hand, a hidden variable model is based on the hypothesis of the existence of a hidden probability measure on the space of "all possible states of the world", such that, each of these possible states is deterministic.

Going back to our experiment, call $p(q, r, s, t)$ to the (hidden) probability that, for a given preparation of the pair of particles, $Q = q$, $R = r$, etc. Then, it is trivial to calculate

$$\begin{aligned} |\mathbb{E}(QS) + \mathbb{E}(RS) + \mathbb{E}(RT) - \mathbb{E}(QT)| &= |\mathbb{E}(QS + RS + RT - QT)| \\ = \Big| \sum_{q,r,s,t} p(q, r, s, t)(qs + rs + rt - qt)\Big| &\leq 2 \sum_{q,r,s,t} p(q, r, s, t) = 2. \end{aligned}$$

This defines an inequality on the set of measurement correlations obtained in the previous experiment,

$$|\mathbb{E}(QS) + \mathbb{E}(RS) + \mathbb{E}(RT) - \mathbb{E}(QT)| \leq 2, \tag{1.1}$$

which is known as *CHSH-inequality.*

Let us assume now that Nature is explained by quantum mechanics and assume that the state formed by both particles is described by

$$|\varphi\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}}.$$

The first qubit goes to Alice, the second to Bob. Alice then measures with the observables $Q = Z$, $R = X$ and Bob measures with observables $S = \frac{-Z-X}{\sqrt{2}}$, $T = \frac{Z_2-X_2}{\sqrt{2}}$ (see Section 2 and Section 3). Then, easy calculations show

$$\langle QS\rangle = \frac{1}{\sqrt{2}},\ \langle RS\rangle = \frac{1}{\sqrt{2}},\ \langle RT\rangle = \frac{1}{\sqrt{2}},\ \langle QT\rangle = -\frac{1}{\sqrt{2}}.$$

Hence,

$$\langle QS\rangle + \langle RS\rangle + \langle RT\rangle - \langle QT\rangle = 2\sqrt{2}. \tag{1.2}$$

Thus, Equation (1.1) and Equation (1.2) tell us that quantum mechanics can not be explained by a local hidden variable model. That is, quantum mechanics predicts that we can obtain certain correlations in the previous measurement-experiment which can not be explained by a local hidden variable model. Nowadays, the verification of the violation of Bell inequalities has become experimental routine (see for instance [**1**] or [**12**]) (albeit there is a remaining desire for a unified loophole-free test).

[1] In fact, for any practical application of this setting one must impose that Alice's choice between measuring property $Q$ or property $R$ is random and similar for Bob.

## 2. Tsirelson's theorem and Grothendieck's theorem

The previous Alice-Bob scenario can be naturally generalized to the case of $N$ measurements. In this case Alice can perform $N$ different measurements $P_1 \cdots, P_N$, each with possible outputs ±1 and similarly to Bob with measurements $Q_1 \cdots, Q_N$. Let us denote

$$\gamma_{i,j} = \mathbb{E}[P_i Q_j], \quad \text{for every} \;\; i,j = 1, \cdots, N.$$

Here, $\mathbb{E}[P_i Q_j]$ denotes the expected value of the product of the outputs of $P_i$ and $Q_j$ for every $i, j$. $\gamma := (\gamma_{i,j})_{i,j=1}^N$ is usually called *correlation matrix.*

Attending the previous section, the correlation matrices obtainable if we assumed a local hidden variable model of Nature are those of the form

$$\gamma_{i,j} = \int_\Omega A_i(\omega) B_j(\omega) d\mathbb{P}(\omega), \tag{2.1}$$

where $(\Omega, \mathbb{P})$ is the hidden probability space and, fixed one of these states $\omega$, $A_i(\omega) = +1$ or $-1$ and similarly for $B_j(\omega)$, for every $i, j$. We call these matrices *classical correlation matrices* and we denote by $\mathcal{L}_N$ the set of classical correlation matrices of size $N$.

On the other hand, according to the postulates of quantum mechanics, in order to define the bipartite system we are measuring on, we must specify a quantum state $\rho \in S_1(\mathbb{C}^n \otimes \mathbb{C}^n)$[2]. On the other hand, each of Alice's two outputs measurements $P_i$ will be described by a POVM $\{E_i, 1 - E_i\}$, where $E_i$ is a positive operator acting on $\mathbb{C}^n$ associated to the output 1 and $1 - E_i$ is a positive operator acting on $\mathbb{C}^n$ associated to the output $-1$. Similarly, we will have to consider the corresponding POVMs to describe Bob's measurements $\{F_j, 1 - F_j\}$ for every $j$. Then, if Alice and Bob perform the measurements $P_i$ and $Q_j$ respectively, we know that the corresponding table of probabilities is given by

$$P(i,j) = \begin{cases} tr\big((E_i \otimes F_j)\rho\big) & \text{is the probability of outputs 1 and 1 respectively} \\ tr\big((E_i \otimes (1 - F_j))\rho\big) & \text{is the probability of outputs 1 and -1 respectively} \\ tr\big(((1 - E_i) \otimes F_j))\rho\big) & \text{is the probability of outputs -1 and 1 respectively} \\ tr\big(((1 - E_i) \otimes (1 - F_j))\rho\big) & \text{is the probability of outputs -1 and -1 respectively.} \end{cases}$$

Then,

$$\begin{aligned} \gamma_{i,j} = \mathbb{E}[P_i Q_j] =& \big[P(1,1|i,j) + P(-1,-1|i,j)\big] - \big[P(-1,1|i,j) + P(1,-1|i,j)\big] \\ =& tr\Big(\big(E_i \otimes F_j + (1 - E_i) \otimes (1 - F_j) - E_i \otimes (1 - F_j) - (1 - E_i) \otimes F_j\big)\rho\Big) \\ =& tr\Big(\big((1 - 2E_i) \otimes (1 - 2F_i)\big)\rho\Big). \end{aligned}$$

Note that if we denote $A_i = 1 - 2E_i$, this is a selfadjoint operators acting on $\mathbb{C}^n$ with $\|A_i\| \leq 1$ for every $i$. On the other hand, every selfadjoint operator $\|A_i\| \leq 1$ can be written as $1 - 2E_i$, where $E_i$ is a positive operator smaller than the identity. Reasoning in a similar way for $B_j = 1 - 2F_j$ for every $j$, we say that $\gamma := (\gamma_{i,j})_{i,j=1}^N$ is a *quantum correlation matrix* if there exit selfadjoint operators $A_1, \cdots, A_N, B_1, \cdots, B_N$ acting on a Hilbert space $\mathbb{C}^n$ with $\max_{i,j=1,\cdots,N}\{\|A_i\|, \|B_j\|\} \leq 1$ and a density operator $\rho$ acting on $\mathbb{C}^n \otimes \mathbb{C}^n$ such that

$$\gamma_{i,j} = tr(A_i \otimes B_j \rho), \quad \text{for every} \;\; i,j = 1, \cdots, N.$$

We denote by $\mathcal{Q}_N$ the set of quantum correlation matrices of order $N$.

It is easy to see that $\mathcal{L}_N \subseteq \mathcal{Q}_N$. Indeed, to see this inclusion let us consider a general element $\gamma \in \mathcal{L}_N$. In fact, for a fixed $N$ we can always assume that the integral in Equation

[2] We assume that Alice's and Bob's systems are described by the same Hilbert space $\mathbb{C}^n$ just for simplicity.

(2.1) is a finite sum. Let us assume that our probability space is of size $K$. Then,

$$\gamma_{i,j} = \sum_{k=1}^{K} p(k) A_i(k) B_j(k),$$

where $A_i(k)$ and $B_j(k)$ are as explained before. Then, considering the $K \times K$ matrices

$$A_i = \begin{pmatrix} A_i(1) & 0 & \cdots & 0 \\ 0 & A_i(2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A_i(K) \end{pmatrix} \text{ and } B_j = \begin{pmatrix} B_j(1) & 0 & \cdots & 0 \\ 0 & B_j(2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & B_j(K) \end{pmatrix}$$

and the $K$-dimensional state $\rho = \sum_{k=1}^{K} p(k)|kk\rangle\langle kk|$, it is trivial to check that

$$tr(A_i \otimes B_j \rho) = \sum_{k=1}^{K} p(k) A_i(k) B_j(k) = \gamma_{i,j}$$

for every $i, j$. Since $A_i$ and $B_j$ are selfadjoint matrices with norm $\leq 1$, the inclusion $\mathcal{L}_N \subseteq \mathcal{Q}_N$ is proved.

One can check that both sets $\mathcal{L}_N$ and $\mathcal{Q}_N$ are convex sets and, moreover, that $\mathcal{L}_N$ is a polytope (it has a finite number of extreme points). Therefore, the set of classical correlation matrices $\mathcal{L}_N$ is described by its facets. The inequalities which describe these facets are usually called *(correlation) Bell inequalities.* Note that one of these inequalities will be of the form

$$\sum_{i,j=1}^{N} M_{i,j}\gamma_{i,j} \leq C \quad \text{for every} \quad \gamma := (\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{L}_N,$$

where $M = (M_{i,j})_{i,j=1}^{N}$ are the coefficients of the corresponding inequality and $C$ is the independent term. Actually, we have already studied one of these inequalities. Indeed, in the case $N = 2$ we have defined the CHSH-inequality in the previous section as that given by

$$M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad C = 2^3.$$

As we showed in the previous section there exist certain quantum correlation matrices $\gamma := (\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{Q}_N$ for which

$$\sum_{i,j=1}^{N} M_{i,j}\gamma_{i,j} = 2\sqrt{2}.$$

In this case, we say that the correlation $\gamma$ violates the corresponding Bell inequality or that we have a *Bell inequality violation.* By convexity, this is equivalent to say that we have a proper content $\mathcal{L}_N \subsetneq \mathcal{Q}_N$.

Note that for every matrix $M = (M_{i,j})_{i,j=1}^{N}$ of real numbers we can associate an inequality

$$|\sum_{i,j=1}^{N} M_{i,j}\gamma_{i,j}| \leq \omega(M), \tag{2.2}$$

[3] In fact, it can be seen that in the case $N = 2$ this is the only Bell inequality up to certain symmetries.

where

$$\begin{aligned}\omega(M) :=& \sup\Big\{\big|\sum_{i,j=1}^{N} M_{i,j}\gamma_{i,j}\big| : (\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{L}_N\Big\}\\ =& \sup\Big\{\big|\sum_{i,j=1}^{N} M_{i,j}t_i s_j\big| : t_i = \pm 1, s_j = \pm 1, \quad \text{for every} \quad i,j\Big\}.\end{aligned}$$

Here, the last equality follows by convexity, since each $(\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{L}_N$ is nothing else than a convex combination of elements of the form $(t_i s_j)_{i,j=1}^{N}$ with $t_i = \pm 1, s_j = \pm 1$ for every $i, j = 1, \cdots, N$. Then, from this point on we will call Bell inequality to any matrix $M = (M_{i,j})_{i,j=1}^{N}$ of real numbers [4] and the value $\omega(M)$ will be called *classical value of* $M$.

On the other hand, we will define the *quantum value of* $M$ by

$$\omega^*(M) := \sup\Big\{\big|\sum_{i,j=1}^{N} M_{i,j}\gamma_{i,j}\big| : (\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{Q}_N\Big\}.$$

The value

$$LV(M) := \frac{\omega^*(M)}{\omega(M)}$$

is usually called *the largest violation of* $M$.

The previously proved content $\mathcal{L}_N \subseteq \mathcal{Q}_N$ means that $\omega^*(M) \geq \omega(M)$ or, equivalently, $LV(M) \geq 1$ for every Bell inequality $M$. Then, $M$ gives rise to a Bell violation whenever $LV(M) > 1$. As a particular example, we have seen that the CHSH-inequality $M_{CHSH}$ verifies

$$LV(M_{CHSH}) \geq \sqrt{2}.$$

In fact, it is not difficult to see that $LV(M_{CHSH}) = \sqrt{2}$. Surprisingly, this value is not far from being optimal, even if we consider matrices of order $N$ as large as we want. This is a consequence of a deep theorem due to Grothendieck in the context of functional analysis. Before going on it, we need to see that the value $\omega^*(M)$ defined above can be expressed in a "much simpler" way.

**2.1. Tsirelson's theorem.** Tsirelson's theorem tells us that, in the same way as the classical value of a Bell inequality $\omega(M)$ can be written as a "combinatorial quantity", the quantum value $\omega^*(M)$ can be understood as a "geometrical quantity".

THEOREM 2.1 (Tsirelson). *Let $\gamma = (\gamma_{i,j})_{i,j=1}^{N}$ be a matrix with real entries. Then, the following statements are equivalent:*

1. $\gamma = (\gamma_{i,j})_{i,j=1}^{N} \in \mathcal{Q}_N$.
2. *There exist norm one elements $x_1, \cdots, x_N, y_1, \cdots, y_N$ in a real Hilbert space such that*

$$\gamma_{i,j} = \langle x_i, y_j\rangle \quad \textit{for every} \quad i,j = 1, \cdots, N.$$

*In particular,*

$$\omega^*(M) := \sup\Big\{\big|\sum_{i,j=1}^{N} M_{i,j}\langle x_i, y_j\rangle\big|\Big\},$$

*where the* sup *is taken over elements $x_1, \cdots, x_N, y_1, \cdots, y_N$ in the unit sphere of a real Hilbert space.*

[4] This is not properly a Bell inequality since the associated inequality (2.2) does not necessarily describe a facet of $\mathcal{L}_N$. However, for the purpose of our study this is completely irrelevant.

In order to prove this theorem, we need to introduce the *Canonical Anticommutation Relations* (CAR)-algebra: Given $N \geq 2$, we consider a set of operators $X_1, \cdots, X_N$, such that they verify the following properties:

1. $X_i^* = X_i$ for every $i = 1, \cdots, N$.
2. $X_iX_j + X_jX_i = 2\delta_{i,j}\mathbb{1}$ for every $i, j \in \{1, \cdots, N\}$.

The proof of the existence of such operators is completely constructive. Indeed, we can construct them as elements of $\bigotimes_{[\frac{n}{2}]} M_2 = M_{2^{[\frac{n}{2}]}}$ by using tensor products of Pauli matrices (see Section 1), where for every positive real number $r$, $[r]$ denotes the least natural number $z$ such that $r \leq z$.

Let us first assume that $N = 2k$ for some $k$. Then, we define the operators:

$$\begin{cases} X_1 = \sigma_x \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1} \otimes \mathbb{1}, & X_2 = \sigma_y \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1} \otimes \mathbb{1} \\ X_3 = \sigma_z \otimes \sigma_x \otimes \cdots \otimes \mathbb{1} \otimes \mathbb{1}, & X_4 = \sigma_z \otimes \sigma_y \otimes \cdots \otimes \mathbb{1} \otimes \mathbb{1} \\ \vdots & \vdots \\ X_{2k-3} = \sigma_z \otimes \sigma_z \otimes \cdots \otimes \sigma_x \otimes \mathbb{1}, & X_{2k-2} = \sigma_z \otimes \sigma_z \otimes \cdots \otimes \sigma_y \otimes \mathbb{1} \\ X_{2k-1} = \sigma_z \otimes \sigma_z \otimes \cdots \otimes \sigma_z \otimes \sigma_x, & X_{2k} = \sigma_z \otimes \sigma_z \otimes \cdots \otimes \sigma_z \otimes \sigma_y. \end{cases}$$

On the other hand, if $N = 2k + 1$, we just add the element

$$X_{2k+1} = \sigma_z \otimes \sigma_z \otimes \cdots \otimes \sigma_z \otimes \sigma_z.$$

Proving that the operators $X_1, \cdots, X_N$ verify the CAR-relations is straightforward.

With these elements at hand, we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. In order to prove the implication 1. ⇒ 2., let us consider the real vector space of self-adjoint operators acting on $H_1 \otimes H_2$, $B(H_1 \otimes H_2)_{sa}$. Then, we can define the Hilbertian space

$$H = (B(H_1 \otimes H_2)_{sa}, \langle \cdot, \cdot \rangle)$$

inherit with the inner product $\langle A, B \rangle = \Re\big(tr(AB\rho)\big)$ for every $A, B \in B(H_1 \otimes H_2)_{sa}$, where $\Re(z)$ denotes the real part of $z$. In fact, by considering the suitable quotient, we can assume this inner product to be definite positive, so that the space $H$ is a real Hilbert space. Furthermore, we can consider the real Hilbert space defined as

$$\tilde{H} = span\{x_i = A_i \otimes \mathbb{1} : i = 1, \cdots, N\} \subset H,$$

and denote by $P : H \to \tilde{H}$ the orthogonal projection onto this space. Then, if we define $y_j = P(\mathbb{1} \otimes B_j)$ for every $j = 1, \cdots, N$, we have a family of elements $x_1, \cdots, x_N, y_1, \cdots, y_N$ in a real Hilbert space of dimension $dim(\tilde{H}) = k \leq N$ verifying $\langle x_i, y_j \rangle = tr(A_i \otimes B_j\rho)$, and such that $\|x_i\| \leq 1, \|y_j\| \leq 1$ for every $i, j = 1, \cdots, N$. Indeed, according to the explanation above

$$\langle x_i, y_j \rangle = \langle A_i \otimes \mathbb{1}, \mathbb{1} \otimes B_j \rangle = tr\big((A_i \otimes \mathbb{1})(\mathbb{1} \otimes B_j)\rho\big) = tr(A_i \otimes B_j\rho)$$

for every $i, j$. Here, we have used that $\Re\big(tr(\Gamma\rho)\big) = tr(\Gamma\rho)$ whenever $\Gamma$ is a self-adjoint operator. On the other hand,

$$\|x_i\| = \langle A_i \otimes \mathbb{1}, A_i \otimes \mathbb{1} \rangle^{\frac{1}{2}} = tr((A_i \otimes \mathbb{1})^2\rho)^{\frac{1}{2}} \leq 1,$$

for every $i$, where we have used that $\|(A_i \otimes \mathbb{1})^2\| = \|A_i\|^2 \leq 1$. A similar argument shows the estimate $\|y_j\| \leq 1$ for every $j$.

Finally, note that we can modify these vectors so that they have norm exactly one, by increasing the dimension of our Hilbert space from $k$ to $k + 2$. Indeed, let us just define

$$\tilde{x}_i = x_i \oplus \sqrt{1 - \|x_i\|^2} \oplus 0 \text{ and } \tilde{y}_j = x_j \oplus 0 \oplus \sqrt{1 - \|y_j\|^2}$$

for every $i, j = 1, \cdots, N$.

To show implication 2. ⇒ 1., let $(\mathbb{R}^M, \langle , \rangle)$ be the real Hilbert space where the norm one elements $(x_i)_{i=1}^N$ and $(y_i)_{j=1}^N$ live. We have seen that we can realize the Clifford operators $X_1, \cdots, X_M$ (of order $M$) as elements in $M_{2^n}$ with $n = [\frac{M}{2}]$. Let us consider the linear map

$$J : \mathbb{R}^M \to CL_M = span\{X_1, \cdots, X_M\}, \text{ defined by } e_k \mapsto X_k \text{ for every } k = 1, \cdots, M.$$

It is very easy to see from the CAR-relations that $\|J : \ell_2^M \to M_{2^n}\| \leq 1$. In particular, for every $x \in \mathbb{R}^M$ with $\|x\| \leq 1$ we have $\|J(x)\| \leq 1$. On the other hand, it is very easy to see that

$$\frac{1}{2^n} tr(J(x)J(y)) = \langle x, y\rangle \text{ for every } x, y \in \mathbb{R}^M.$$

In fact, if we consider the state $|\psi\rangle = \frac{1}{2^{\frac{n}{2}}} \sum_{i,j=1}^{2^n} |ij\rangle \in \mathbb{C}^{2^n} \otimes_2 \mathbb{C}^{2^n}$, it is straightforward to check that for every $A, B \in M_{2^n}$ we have

$$\frac{1}{2^n} tr(AB) = tr(A \otimes B|\psi\rangle\langle\psi|) = \langle\psi|A \otimes B|\psi\rangle.$$

Therefore, if we define the operators $A_i = J(x_i) \in M_{2^n}, B_j = J(y_j) \in M_{2^n}$ for every $i, j$, we obtain a family of self adjoint operators with norm lower than or equal to one, and such that

$$\langle\psi|A_i \otimes B_j|\psi\rangle = \frac{1}{2^n} tr(A_i B_j) = \frac{1}{2^n} tr(J(x_i)J(y_j)) = \langle x, y\rangle = \gamma_{i,j}$$

for every $i, j = 1, \cdots, N$. This concludes the proof. □

**2.2. Grothendieck's theorem.** Theorem 2.1 allows us to bring Grothendieck's *fundamental theorem in the metric theory of tensor products* to the context of Bell inequalities. This last result, proved by Grothendieck in the context of tensor products of Banach space ([**5**]), is a central result in Banach space theory. Among many equivalent reformulations of this result one can find

THEOREM 2.2 (Grothendieck's inequality). *There exits a positive universal constant $K_G$ such that for every natural number $N$ and for every matrix of real coefficients $(M_{i,j})_{i,j=1}^N$ the following inequality holds:*

$$\sup\Big\{\big|\sum_{i,j=1}^N M_{i,j}\langle x_i, y_j\rangle\big| : \|x_i\|, \|y_j\| = 1 \ \forall i, j\Big\} \leq K_G \cdot \sup\Big\{\big|\sum_{i,j=1}^N M_{i,j} t_i s_j\big| : t_i, s_j = \pm 1 \ \forall i, j\Big\}.$$

*Here, the first supremum is taken over families of vectors $x_1, \cdots, x_N, y_1, \cdots, y_N$ in an arbitrary real Hilbert space.*

The constant $K_G$, known as *the (real) Grothendieck's constant*, verifies

$$1.67696... \leq K_G < \frac{\pi}{2\log((1 + \sqrt{2})} = 1.7822139781...$$

Howevere, the exact value of this constant is still an open question.

According to our previous description of the values $\omega(M)$ and $\omega^*(M)$, Theorem 2.2 can be reformulated as

THEOREM 2.3. *There exits a positive universal constant $K_G$ such that for every natural number $N$ and for every Bell inequality $(M_{i,j})_{i,j=1}^N$ the following inequality holds:*

$$\omega^*(M) \leq K_G \cdot \omega(M) \quad \textit{or, equivalently, } LV(M) \leq K_G.$$

Therefore, the basic example in $N = 2$ defined by the CHSH-inequality, $M_{CHSH}$, which provides a Bell violation $LV(M_{CHSH}) \geq \sqrt{2}$, is "close to be optimal".

As we said in the introduction of this chapter, quantum nonlocality has become a crucial point in many different areas of quantum information like quantum cryptography and communication complexity. The basic idea is that quantum correlations which are not classical, so those which violate a Bell inequality, can be used to define "certain protocols" with some advantages over those protocols defined according to the classical theory. That is, the context of Bell inequalities (or quantum nonlocality) allows us to "realize the advantages of quantum mechanics with respect to the classical theory". Furthermore, the amount of Bell violation $LV(M)$ is a quantifier of how better quantum mechanics is with respect to classical mechanics. Therefore, Theorem 2.3 must be understood as a limitation of quantum mechanics. This motivated Tsirelson to ask in [**14**] whether one could get larger violations if we consider a more general context. In particular, the previous study can be done in a completely analogous way if one considers three people: Alice, Bob and Charlie, in the measurement process. Indeed, in this case we will obtain correlations $\gamma = (\gamma_{i,j,k})_{i,j,k=1}^N$ which will be called *classical* if

$$\gamma_{i,j,k} = \int_\Omega A_i(\omega)B_j(\omega)C_k(\omega)d\mathbb{P}(\omega),$$

where $(\Omega, \mathbb{P})$ is the hidden probability space and, fixed one of these states $\omega$, $A_i(\omega) = +1$ or $-1$ and similarly for $B_j(\omega)$ and $C_k(\omega)$, for every $i, j, k$. On the other hand, $\gamma$ will be called a *quantum correlation* if there exit selfadjoint operators $A_1, \cdots, A_N$, $B_1, \cdots, B_N$, $C_1, \cdots, C_N$ acting on a Hilbert space $\mathbb{C}^n$ with $\max_{i,j,k=1,\cdots,N}\{\|A_i\|, \|B_j\|, \|C_j\|\} \leq 1$ and a density operator $\rho$ acting on $\mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n$ such that

$$\gamma_{i,j,k} = tr(A_i \otimes B_j \otimes C_k\rho), \quad \text{for every} \quad i, j, k = 1, \cdots, N.$$

In the same way as before, for a given Bell inequality $M = (M_{i,j,k})_{i,j,k=1}^N$ we can define its *largest violation* by

$$LV(M) := \frac{\omega^*(M)}{\omega(M)},$$

where $\omega(M)$ and $\omega^*(M)$ are defined as in the previous section.

It turns out that Tsirelson's problem has a positive answer! In [**10**] (see also [**3**]) the following result was proved.

Theorem 2.4. *For every positive real number $D$, there exists a high enough natural number $N$ and a Bell inequality $M = (M_{i,j,k})_{i,j,k=1}^N$ such that*

$$LV(M) \geq D.$$

Notice the difference between Theorem 2.3 and Theorem 2.4. In this last result, an unlimited amount of violation can be obtained if we allow the size $N$ of the tensor (matrix) $M$ to increase enough. The best estimate so far was proved in [**3**], where the authors showed that $N \simeq D^4$ suffices. On the other hand, this unlimited amount of violation is not possible in the bipartite case, where $K_G$ is an upper bound for $LV(M)$ for every $M$ (independently of its size $N$). Hence, in the tripartite scenario we could, in principle, obtain unlimited advantages by using quantum mechanics rather than classical resources.

CHAPTER 6

# Some notions about classical information theory

In this chapter we will deal with two important questions in (classical) information theory:

1. How much can a message be “compressed”? (Noiseless coding theorem).
2. Which is the best rate of reliable communication through a noisy channel? (Noisy channel coding theorem).

## 1. Shannon’s noiseless channel coding theorem

Shannon’s noiseless channel coding theorem quantifies how much we can compress the information being produced by a classical information source. A simple but very fruitful model of a classical information source consists of a sequence of random variables $X_1, X_2, \cdots$ whose values represent the outputs of the source. We will assume that the random variables take values from a finite alphabet of symbols. Furthermore, we assume that the different uses of the source are *independent* and *identically distributed*; that is, the source is what is known as an i.d.d. information source. Let us assume that our alphabet has $k$ letters $\{a_1, \cdots, a_k\}$ so that the probability distribution of the random variable $X$ is given by $\big(p_x = p(a_x)\big)_{x=1}^k$. A classical example is a binary alphabet $\{0,1\}$ so that $p(0) = 1-p$ and $p(1) = p$.

Given a probability distribution $(p_x)_{x=1}^k$ associated to a random variable $X$, we define the *Shannon entropy* of $X$ by

$$H(X) = \sum_{x=1}^k p_x(-\log p_x).$$

It is easy to see that this quantity is a concave function[1] verifying

$$0 \leq H(X) \leq \log k, \tag{1.1}$$

for every probability distribution $(p_x)_{x=1}^k$.

Indeed, the first inequality is trivial because $H(X)$ is defined as a sum of positive numbers. For the second inequality we will make use of the basic inequality

$$x - 1 \geq \ln x = \ln 2 \log x \quad \text{for every} \quad x \geq 0. \tag{1.2}$$

With this inequality at hand, we have

$$\begin{aligned} \log k - H(X) &= \sum_x p_x\big(\log k + \log p_x\big) = \sum_x p_x(-\log \frac{1}{kp_x}) \\ &\geq \frac{1}{\ln 2}\sum_x p_x(1 - \frac{1}{kp_x}) = \frac{1}{\ln 2}\sum_x (p_x - \frac{1}{k}) = 0. \end{aligned}$$

In fact, the previous argument shows that $H(X) = \log k$ if and only if $p_x = \frac{1}{k}$ for every $k$.

Let us now consider long messages

$$x_1 \cdot \cdots \cdot x_n$$

[1] We will see this for the von Neumman entropy, which implies the same property for the Shannon entropy.

with $n$ letters ($n \gg 1$) and we ask: Is it possible to compress the message to a shorter string of letters that convey essentially to the same information? By independence we have

$$p(x_1 \cdot \dots \cdot x_n) = p(x_1) \cdot \dots \cdot p(x_n).$$

For every $\epsilon > 0$ we say that the string of source symbols $x_1 \cdot \dots \cdot x_n$ is *$\epsilon$-typical* if

$$2^{-n\big(H(X)+\epsilon\big)} \leq p(x_1 \cdot \dots \cdot x_n) \leq 2^{-n\big(H(X)-\epsilon\big)}$$

and denote the set of all $\epsilon$-typical sequences of length $n$ by $T(n, \epsilon)$. Note that an equivalent formulation of the condition above is

$$\Big|\frac{1}{n} \log \frac{1}{p(x_1 \cdot \dots \cdot x_n)} - H(X)\Big| \leq \epsilon.$$

THEOREM 1.1 (Theorem of typical sequences).

1. *Fix $\epsilon > 0$. Then, for any $\delta > 0$, for sufficiently large $n$, the probability that a sequence is $\epsilon$-typical is at least $1 - \delta$.*
2. *For any fixed $\epsilon > 0$ and $\delta > 0$, for sufficiently large $n$, the number $|T(n, \epsilon)|$ of $\epsilon$-typical sequences satisfies*
$$(1 - \delta)2^{n\big(H(X)-\epsilon\big)} \leq |T(n, \epsilon)| \leq 2^{n\big(H(X)+\epsilon\big)}.$$
3. *Let $S(n)$ be a collection of size of at most $2^{nR}$ of length $n$ sequences from the source, where $R < H(X)$ is fixed. Then, for any $\delta > 0$ and sufficiently large $n$, $\sum_{x \in S(n)} p_x \leq \delta$.*

PROOF. We first note that $-\log p(X_i)$ are independent and identically distributed random variables. By the law of large numbers we know that for every $\epsilon > 0$ and $\delta > 0$, for sufficient large $n$ we have

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} - \log p(X_i) - \mathbb{E}[-\log p(X)]\Big| > \epsilon\Big) < \delta.$$

Now, $\mathbb{E}[-\log p(X)] = H(X)$ and

$$-\sum_{i=1}^{n} \log p(X_i) = -\log\Big(p(X_1)\cdots p(X_n)\Big) = -\log\Big(p(X_1 \cdots X_n)\Big).$$

Thus,

$$P\Big(\Big|\frac{1}{n}\log\frac{1}{p(X_1 \cdot \dots \cdot X_n)} - H(X)\Big| \leq \epsilon\Big) \geq 1 - \delta.$$

In order to prove the second assertion, note that by the first part of the theorem we know that

$$1 \geq \sum_{x \in T(n,\epsilon)} p(x) \geq \sum_{x \in T(n,\epsilon)} 2^{-n\big(H(X)+\epsilon\big)} = |T(n, \epsilon)| 2^{-n\big(H(X)+\epsilon\big)},$$

from which we deduce that $|T(n, \epsilon)| \leq 2^{n\big(H(X)+\epsilon\big)}$.

On the other hand,

$$1 - \delta \leq \sum_{x \in T(n,\epsilon)} p(x) \leq \sum_{x \in T(n,\epsilon)} 2^{-n\big(H(X)-\epsilon\big)} = |T(n, \epsilon)| 2^{-n\big(H(X)-\epsilon\big)},$$

from which we deduce that $|T(n, \epsilon)| \geq (1 - \delta)2^{n\big(H(X)-\epsilon\big)}$.

Finally, in order to show the third part of the theorem we choose $\epsilon > 0$ so that $R < H(X) - \delta$ and $0 < \epsilon \leq \frac{\delta}{2}$. Then, we split the sequences in $S(n)$ up into the $\epsilon$-typical and the $\epsilon$-atypical sequences. According to 1., for sufficiently large $n$ the total probability of atypical sequences is $\leq \frac{\delta}{2}$. There are at most $2^{nR}$ typical sequence in $S(n)$, each with probability at most $2^{-n\big(H(X)-\epsilon\big)}$, so the probability of the typical sequences is at most $2^{-n\left(H(X)-\epsilon-R\right)}$, which goes to zero as $n$ goes to infinity. Thus, the total probability of the sequences in $S(n)$ is less than $\delta$ for $n$ sufficiently large. □

Let $X_1, X_2, \cdots$ be an i.i.d. classical information source over a finite alphabet containing $k$ symbols. A *compression scheme of rate* $R$ maps possible sequences $x = (x_1 x_2 \cdots x_n)$ to a bit string of length $nR$ which we denote by $C^n(x) = C^n(x_1 x_2 \cdots x_n)$ (we understand $nR = [nR]$, the smallest natural number $p$ such that $nR \leq p$). The matching *decompression scheme* takes the $nR$ compressed bits and maps them back to string of $n$ letters from the alphabet, $D^n(C^n(x))$. A compression-decompression scheme $(C^n, D^n)$ is said to be reliable if the probability that $D^n(C^n(x)) = x$ approaches one as $n$ tends to infinity.

THEOREM 1.2 (Shannon's noiseless channel coding theorem). *Suppose $\{X_i\}$ is an i.i.d. information source with entropy rate $H(X)$. Suppose $R > H(X)$. Then, there exists a reliable compression scheme of rate $R$ for the source. Conversely, if $R < H(X)$, then any compression scheme of rate $R$ will not be reliable.*

PROOF. Suppose $R > H(X)$. Choose $\epsilon > 0$ such that $H(X) + \epsilon < R$. Consider the set $T(n, \epsilon)$ of $\epsilon$-typical sequences. For any $\delta > 0$ and for sufficiently large $n$, there are at most $2^{n\left(H(x)+\epsilon\right)} < 2^{nR}$ such sequences, and the probability of the source producing such a sequence is at least $1 - \delta$. Since the number of typical sequences is lower than $2^{nR}$, we can consider an enumeration of them by using $nR$ bits. The method compression $C^n$ is simply to check the output of the source to see if it is $\epsilon$-typical. If it is not, then compress to some fixed $nR$-bit string (we don't care which one). If the output of the source is typical then we compress the output simply by associating it with its corresponding $nR$-bit string according to our enumeration. Since this is an enumeration, we can always recover back our message. Indeed, the decompression scheme $D^n$ will send each of the $nR$-bit strings involved in our enumeration to its corresponding message and the rest of the $nR$-bit strings (those which have not been used in our enumeration) to a fixed message (we don't care which one). Then, we see that for any typical sequence $x = x_1 x_2 \cdots x_n$ we have $D^n(C^n(x)) = x$. Since the probability of the source produc ing such a sequence is at least $1 - \delta$ (for arbitrarily small $\delta$) we have the required reliability condition.

Suppose that $R < H(X)$. The combined compression-decompression operation has at most $2^{nR}$ possible outputs, so at most $2^{nR}$ of the sequences output from the source can be compressed and decompressed without an error occurring. By the theorem of typical sequences, for sufficiently large $n$ the probability of a sequence output from the source lying in a subset of $2^{nR}$ sequences goes to zero, for $R < H(X)$. Thus, any such compression scheme can not be reliable. □

The previous result provides the Shannon's entropy with a clear information theoretical meaning: $H(X)$ quantifies how much information is conveyed, on the average, by a letter drawn from the ensemble $X$, for it tells us how many bits are required (asymptotically as $n$ goes to infinity, when $n$ is the number of letters drawn) to encode that information.

## 2. Conditional entropy and Fano's inequality

In order to study the noisy channel coding theorem, we need to consider first the situation where we have two random variables $X$ and $Y$. Let us denote by $(X, Y)$ the joint system with joint probability distribution $(p_{x,y})_{x,y=1}^{k}$[2]. We can consider the *joint Shannon entropy* by

[2] We assume the same number of letters in the alphabets for $X$ and $Y$ just to simplify notation.

naturally defining

$$H(X,Y) = -\sum_{x,y} p_{x,y} \log p_{x,y}.$$

Note that $H(X,Y) = H(Y,X)$. However, in this case we can also measure how uncertainty we are, on average, about the value of $X$, given that we know the value of $Y$. To this end, we define the *Shannon entropy of $X$ conditioned on knowing a particular value of $y = y_0$* as

$$H(X|Y = y_0) = -\sum_{x} p(x|y = y_0) \log p(x|y = y_0).$$

This entropy quantifies our uncertainty about $X$ when we know the value $y = y_0$. On the other hand, a more relevant quantity is obtained when we calculate the average, in $Y$, of the previous quantity. We define the *entropy of $X$ conditioned on knowing the value of $Y$* by

$$H(X|Y) = \sum_{y} p(y) H(X|Y = y) = -\sum_{x,y} p_{x,y} \log p(x|y) = H(X,Y) - H(Y).$$

Obviously this is a measure of how uncertainty we are about $X$ when we know the value of $Y$. A crucial quantity for us will be the mutual information content of $X$ and $Y$, which measures how much information $X$ and $Y$ have in common. The information about $X$ that we gain when we learn $Y$ is quantified by how much the number of bits per letter needed to specify $X$ is reduced when $Y$ is known. Thus, the *mutual information of $X$ and $Y$* is defined as

$$\begin{aligned} H(X:Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) = H(Y:X). \end{aligned}$$

Note that this quantity can also be understood in the following sense. Suppose we add the information content of $X$, $H(X)$, to the information content of $Y$. Information which is common to $X$ and $Y$ will have been counted twice in this sum, while information which is not common will have been counted exactly once. Substracting off the joint information of $(X,Y)$, $H(X,Y)$, we therefore obtain the common information of $X$ and $Y$.

We next summarize some of the most important properties of the Shannon entropy.

1. $H(X|Y) \geq 0$. Therefore,
   1.1. $H(X,Y) \geq H(Y)$ (resp. $H(X,Y) \geq H(X)$) with equality if and only if $X = f(Y)$ (resp. $Y = f(X)$).
   1.2. $H(X:Y) \leq H(Y)$ (resp. $H(X:Y) \leq H(X)$) with equality if and only if $X = f(Y)$ (resp. $Y = f(X)$).
   The proof of this result is trivial from the very definition of $H(X|Y)$.

3. *Subadditivity*: $H(X,Y) \leq H(X) + H(Y)$ with equality if and only if $X$ and $Y$ are independent random variables.
   We will prove this property for the von Neumann entropy, which implies in particular the same result for the Shannon entropy.

4. $H(X|Y) \leq H(X)$ and thus $H(X:Y) \geq 0$ with equality in each if and only if $X$ and $Y$ are independent random variables.
   The proof of this result follows from subadditivity and the relevant definitions.

5. *Chaining rule for conditional entropies*: Let $X_1, \cdots, X_n$ and $Y$ be any set of random variables. Then,

$$H(X_1, \cdots, X_n|Y) = \sum_{i=1}^{n} H(X_i|Y, X_1, \cdots, X_{i-1}).$$

Note that the case $n = 2$ follows by definition and linear algebra

$$\begin{aligned} H(X_1, X_2|Y) &= H(X_1, X_2, Y) - H(Y) \\ &= H(X_1, X_2, Y) - H(X_1, Y) + H(X_1, Y) - H(Y) \\ &= H(X_2|Y, X_1) + H(X_1|Y). \end{aligned}$$

The general case $n$ follows easily by an inductive argument.

**2.1. Fano's inequality.** Suppose we want to infer the value of a random variable $X$ based on our knowledge about a random variable $Y$. The quantity $H(X|Y)$ should be a good measure of how difficult to do this is. This intuition is quantified by the Fano's inequality.

THEOREM 2.1 (Fano's inequality). *Let $\tilde{X} = f(Y)$ be a function of the random variable $Y$ which we are using as our best guess for the value of the random variable $X$. Let $p_e = p(X \neq \tilde{X})$ be the probability that this guess is not correct. Then,*

$$H_b(p_e) + p_e \log(|X| - 1) \geq H(X|Y),$$

*where $H_b(\cdot)$ is the binary entropy and $|X|$ is the number of values that $X$ can assume.*

PROOF. Let us define the error random variable

$$\begin{cases} E \equiv 1 & \text{if} \quad X \neq \tilde{X} \\ E \equiv 0 & \text{if} \quad X = \tilde{X}. \end{cases}$$

Note that

1. $H(E) = H(p_e)$,
2. $H(E|X, Y) = 0$, and
3. $H(E|Y) \leq H(E) = H(p_e)$.

Then, applying the chaining rule twice we obtain

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

and

$$H(E, X|Y) = H(E|Y) + H(X|E, Y) \leq H(p_e) + H(X|E, Y).$$

Therefore,

$$H(X|Y) \leq H(p_e) + H(X|E, Y).$$

The proof of Fano's inequality is reduced then to upper bound $H(X|E, Y)$. Now,

$$\begin{aligned} H(X|E, Y) &= p(E = 0)H(X|E = 0, Y) + p(E = 1)H(X|E = 1, Y) \\ &\leq p(E = 0)\dot{0} + p_e \log(|X| - 1) \\ &= p_e \log(|X| - 1). \end{aligned}$$

Here, we have used that when $E = 1$, $X \neq f(Y)$ and $X$ can assume at least $|X| - 1$ values. Therefore, $H(X|E = 1, Y) \leq \log(|X| - 1)$ follows from Equation (1.2).

The proof is then concluded. □

## 3. Shannon's noisy channel coding theorem: Random coding

As we explained in Section 4, a classical channel is given by a stochastic action denoted by $\mathcal{N}:\ell_1^n\to\ell_1^n$. The *capacity of a channel* is defined as an asymptotic ratio:

$$\frac{\text{number of transmitted bits with an } \epsilon\to 0 \text{ error}}{\text{number of required uses of the channel in parallel}}.$$

More precisely, given a channel $\mathcal{N}:\ell_1^n\to\ell_1^n$, its capacity is defined as

$$C_c(\mathcal{N}):=\lim_{\epsilon\to 0}\limsup_{k\to\infty}\Big\{\frac{m}{k}:\exists_{\mathcal{A}}\exists_{\mathcal{B}}\text{ such that }\|id_{\ell_1^{2^m}}-\mathcal{B}\circ\mathcal{N}^{\otimes_k}\circ\mathcal{A}\|<\epsilon\Big\},$$

where the encoder $\mathcal{A}:\ell_1^{2^m}\to\bigotimes^k\ell_1^n$ and decoder $\mathcal{B}:\bigotimes^k\ell_1^n\to\ell_1^{2^m}$ are channels and $\mathcal{N}^{\otimes_k}:\bigotimes^k\ell_1^n\to\bigotimes^k\ell_1^n$ denotes the use of $k$ times the channel in parallel. The composition $\mathcal{B}\circ\mathcal{N}^{\otimes_k}\circ\mathcal{A}$ represents the protocol in which Alice encodes a message, sends this information to Bob using $k$ times the channel in parallel and Bob decodes the information that he receives.

Shannon's noisy channel coding theorem provides an elegant formula for the capacity of a channel as an optimization function over the mutual information between the input distributions $(p(x))_x$ for $X$, for one use of the channel, and the corresponding induced random variable $Y$ at the output of the channel $(\mathcal{N}(P))_y$ (see Theorem 3.1). We will not prove Shannon's noisy channel coding theorem in detail here because we will show a more general result in the next chapter. Indeed, the Holevo-Schumacher-Westmoreland's Theorem (Theorem 4.1) gives a formula for the classical capacity of a quantum channel (when the encoding is restricted to the use of product states) and, in particular, this result generalizes Shannon's theorem for classical channels. However, it is very interesting to understand the idea of the proof in the classical context so that we gain some intuition about how to proceed in the quantum case. In fact, we will start by focusing on the particular binary symmetric channel introduced in Chapter 4, Section 4. This channel $\mathcal{N}$ acts on single bits ($n=2$ above) and it is described by the following stochastic action $\big(P(x|y)\big)_{x,y=0,1}$

$$\begin{cases} P(0|0)=1-p, & P(0|1)=p;\\ P(1|0)=p, & P(1|1)=1-p.\end{cases}$$

Note that for every input string of length $k$ bits $x_1x_2\cdots x_k$, errors (due to $k$ uses of the channel in parallel) will typically cause about $kp$ of these bits to flip. Hence, the input will be typically distorted to one of about $2^{kH(p)}$ output strings. This defines a sphere of Hamming radius $kp$ around the input string. Indeed, given a fixed string, the number of strings obtained from the first one by flipping $kp$ bits is given by $\binom{k}{kp}$. Then, according to Stirling approximation formula $\log k!=k\log k-k+0(\log k)$, we obtain

$$\begin{aligned}\log\binom{k}{kp}&=\log\Big(\frac{k!}{(kp)![k(1-p)]!}\Big)\\ &\simeq k\log k-k-\big(kp\log kp-kp+k(1-p)\log k(1-p)-k(1-p)\big)\\ &=k\big(-p\log p-(1-p)\log(1-p)\big)=kH(p),\end{aligned}$$

where $H(p)=-p\log p-(1-p)\log(1-p)$ is the Shannon entropy of the binary distribution $\{p,1-p\}$.

Let us assume that the capacity of this channel is $R$. That is, we can send $kR$ bits (so $2^{kR}$ codewords) with the use of $k$ times the channel in parallel. To decode reliably, we want to choose our input codewords so that the error sphere of two different codewords are unlikely to overlap. Otherwise, two different inputs will sometimes yield the same output, and decoding errors will inevitably occur. If we want to avoid these ambiguities, the total number of strings contained in all $2^{kR}$ error spheres must not exceed the total number of bits in the input message $2^k$. Then, we require

$$2^{kH(p)}2^{kR} \leq 2^k, \quad \text{or} \quad R \leq 1 - H(p) \equiv C(p).$$

In the following, we will "show" that such an upper bound can be attained asymptotically. The basic idea of Shannon is that $C(p)$ can be attained by considering an average on *random codes*. This is somehow surprising since this does not seem the most clever way to choose a code. However, we will see that this procedure is optimal!

Let us fixed an ensemble $X = \{q, 1-q\}$ which will describe a particular classical source with associated alphabet $\{0,1\}$. Suppose that $2^{kR}$ codewords are chosen at random by sampling the ensemble $X^k$. A message (one of the codewords) is sent to Bob. To decode this message, Bob will consider a "Hamming sphere" around the *message received* that contains $2^{k(H(p)+\delta)}$ strings. The message is decoded as the codeword contained in the sphere, assuming such a codeword exists and is unique. Otherwise, we will assume that a decoding error occurs.

How likely is a decoding error? We have chosen the decoding sphere large enough so that failure of a valid codeword to appear in the sphere is atypical, so we only need to worry about the existence of more than one valid codeword in the same sphere. Since there are all together roughly $2^{kH(q)}$ possible strings, the Hamming sphere around the output contains a fraction

$$\frac{2^{k(H(p)+\delta)}}{2^{kH(q)}} = 2^{-k(H(q)-H(p)-\delta)}$$

of all strings. Thus, the probability that one of the $2^{kR}$ randomly chosen codewords occupies this sphere "by accident" is

$$2^{-k(H(q)-H(p)-\delta)}2^{kR} = 2^{-k(H(q)-H(p)-R-\delta)}.$$

We can understand the previous quantity as the expected value (in the codes) of the average error value (for a fixed code). Indeed, the symmetry of our argument implies

$$\mathbb{E}_C\big[\frac{1}{2^{kR}}\sum_{i=1}^{2^{kR}} P_i^C\big] = \frac{1}{2^{kR}}\sum_{i=1}^{2^{kR}} \mathbb{E}_C[P_i^C] = \frac{1}{2^{kR}}\sum_{i=1}^{2^{kR}} \mathbb{E}_C[P_1^C] = \mathbb{E}_C[P_i^C],$$

where $\mathbb{E}_C$ denotes the average in the codes and, for a fixed code $C$, $P_i^C$ denotes the probability of a decoding error for the codeword $i \in C$. Since we may chose $\delta$ as small as we want, $R$ can be chosen as close to $H(q) - H(p)$ as we want (but below $H(q) - H(p)$) and this error probabilities will still become exponentially small as $k \to \infty$. Since this happens for an arbitrary ensemble $X = \{q, 1-q\}$, we can maximize over them to obtain $H(q) = 1$ for $q = \frac{1}{2}$. Therefore, we can achieve a capacity $1 - H(p) = C(p)$.

So far, we have shown that "the average" probability of error is small, where we average over the choice of random code, and for each specific code, we also average over all codewords. Thus, there must exist one particular code with average probability of error (average over the codewords) less than $\epsilon$. But we want to have that the probability of error is small for every codeword. To establish this stronger result, let us denote by $P_i$ the probability of a decoding error when codeword $i$ is sent. We know that

$$P_{av} = \frac{1}{2^{kR}}\sum_{i=1}^{2^{kR}} P_i < \epsilon.$$

Let $N_{2\epsilon}$ denote the number of codewords with $P_i > 2\epsilon$. Then, we must have

$$N_{2\epsilon} < 2^{kR-1}.$$

Therefore, we can redefine our new code, formed by those codewords verifying $P_i < 2\epsilon$ to obtain a new code with probability of error $< 2\epsilon$ for every $i$ and rate

$$Rate = R - \frac{1}{k} \to R \ (k \to \infty).$$

Therefore, we have shown that $C(p) = 1 - H(p)$ is the capacity of the symmetric binary channel.

The previous argument can be "easily" generalized to arbitrary channels $\mathcal{N} = \big(P(x|y)\big)_{x,y}$ to show

THEOREM 3.1 (Noisy channel coding theorem). *For a noisy channel $\mathcal{N} : \ell_1^n \to \ell_1^n$ the capacity is given by*

$$C_c(\mathcal{N}) = \max_{P=(p(x))_x} H(X:Y),$$

*where the maximum is taken over all input distributions $(p(x))_{x=1}^n$ for $X$, for one use of the channel, and $Y$ is the corresponding induced random variable at the output of the channel $(\mathcal{N}(P))_{y=1}^n$.*

Let us start by explaining how to obtain the estimate

$$C_c(\mathcal{N}) \geq \max_{P=(p(x))_x} H(X:Y). \tag{3.1}$$

Let us take $X = \{x, p_x\}$ an arbitrary probability distribution for the input letters. Once we know $X$ and $\mathcal{N} = \big(P(x|y)\big)_{x,y}$, we can determine $Y = \{y, p_y\}$. Again, we will consider an average over random codes, where codewords are chosen with a priori probability governed by $X^k$. Thus, with high probability , these codewords will be chosen from a typical set of strings of letters, where there are about $2^{kH(X)}$ such typical strings. For a typical received message in $Y^k$, there are about $2^{kH(X|Y)}$ messages that could have been sent. We may decode by associating with the received a "sphere" containing $2^{k\big(H(X|Y)+\delta\big)}$ possible inputs. If there exists a unique codeword in this sphere, we decode the message as that codeword.

As before, it is unlikely that no codewords will be in the sphere, but we must exclude the possibility that there are more than one. Each decoding sphere contains a fraction

$$\frac{2^{k\big(H(X|Y)+\delta\big)}}{2^{kH(X)}} = 2^{-k\big(H(X)-H(X|Y)-\delta\big)} = 2^{-k\big(H(X:Y)-\delta\big)}$$

typical inputs. If there are $2^{kR}$ codewords, the probability that any one falls in the decoding sphere by accident is

$$2^{kR}2^{-k\big(H(X:Y)-\delta\big)} = 2^{-k\big(H(X:Y)-R-\delta\big)}.$$

Since $\delta$ can be chosen arbitrarily small, we can chose $R$ as close to $H(X:Y)$ as we want (but smaller than $H(X:Y)$) and still works.

As in the case of the binary symmetric channel, the previous quantity must be understood as the expected value (in the codes) of the average error for a fixed code. Therefore, the previous argument shows that we have the result on average. A similar argument as the one explained before for the binary symmetric channel allows us to obtain (3.1). In order to prove inequality

$$C_c(\mathcal{N}) \leq \max_{P=(p(x))_x} H(X:Y). \tag{3.2}$$

let us assume that we have $2^{kR}$ strings of $k$ letters as our codewords. Let us consider a probability distribution (denoted by $\tilde{X}^k$) in which each codeword occurs with probability $2^{-kR}$. Note that we have

$$H(\tilde{X}^k) = kR.$$

Sending the codewords through the channel we obtain a probability distribution $\tilde{Y}^k$ of output strings. Since we assume that the channel acts on each letter independently, the conditional probability for string of $k$ letters factorizes as

$$p(y_1y_2\cdots y_k|x_1x_2\cdots x_k) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_k|x_k).$$

Then, just by considering the definition of the conditional entropy we obtain that

$$H(\tilde{Y}^k|\tilde{X}^k) = \sum_{i=1}^{k} H(\tilde{Y}_i|\tilde{X}_i),$$

where $\tilde{X}_i$ and $\tilde{Y}_i$ are the marginal probability distributions for the $i^{th}$ letter determined by our distribution on the codewords. According to the subadditivity of the Shannon entropy we have

$$H(\tilde{Y}^k) \leq \sum_{i=1}^{k} H(\tilde{Y}_i).$$

Therefore,

$$\begin{aligned} H(\tilde{Y}^k : \tilde{X}^k) = H(\tilde{Y}^k) - H(\tilde{Y}^k|\tilde{X}^k) &\leq \sum_{i=1}^{k} \big(H(\tilde{Y}_i) - H(\tilde{Y}_i|\tilde{X}_i)\big) \\ &\leq \sum_{i=1}^{k} H(\tilde{Y}_i : \tilde{X}_i) \leq k \max_{P=(p(x))_x} H(X:Y). \end{aligned}$$

Let us denote by $\overline{X}^k = f(\tilde{Y}^k)$ the result of Bob's decoding from the random variable $\tilde{Y}^k$. We see that

$$p_{av} := p(\tilde{X}^k \neq \overline{X}^k) = \frac{1}{2^{kR}} \sum_{x \text{ codewords}} p(x \neq \overline{x}),$$

where here $\overline{x}$ denotes the result of Bob's decoding when Alice has sent the codeword $x$ through the $k$ uses of the channel. Note that the previous quantity $p_{av}$ is exactly the average error in the protocol performed by Alice and Bob. Our goal is to show that this probability is bounded away from zero if $R > \max_{P=(p(x))_x} H(X:Y)$. Then, the maximum error probability will be also bounded away from. Indeed, by Fano's inequality, we have

$$H_b(p_{av}) + p_{av}kR \geq H(\tilde{X}^k|\tilde{Y}^k).$$

Then, we conclude that

$$\begin{aligned} p_{av}kR \geq H(\tilde{X}^k|\tilde{Y}^k) - H_b(p_{av}) &= H(\tilde{X}^k) - H(\tilde{X}^k : \tilde{Y}^k) - H_b(p_{av}) \\ &\geq kR - k \max_{P=(p(x))_x} H(X:Y) - H_b(p_{av}). \end{aligned}$$

Hence,

$$p_{av} \geq \frac{1}{R}\Big(R - \max_{P=(p(x))_x} H(X:Y) - \frac{H_b(p_{av})}{k}\Big).$$

That is, if $R > \max_{P=(p(x))_x} H(X:Y)$ we obtain a positive lower bound for $p_{av}$ independent of $k$. This concludes the proof.

**3.1. A remark on the encoder.** There is something to remark about the proof of Theorem 3.1. In the previous upper bound (3.2) for the classical capacity of a channel, we have restricted to those encoders $\mathcal{A} : \ell_1^{2^{kR}} \to \otimes^k \ell_1^n$ in which each $|i\rangle \in \ell_1^{2^{kR}}$ is sent to an element $|i_1 \cdots i_k\rangle \in \otimes^k \ell_1^n$, where $|i_j\rangle \in \ell_1^n$ is an element of the computational basis. According to the notation in the previous proof, this means that each codeword $\overline{x}$ is of the form $x_1 \cdots x_k$, where each $x_i$ is a symbol in our alphabet. However, more general encoders could be used, namely those which send each $|i\rangle$ to a general element $w_i \in \otimes^k \ell_1^n$. In other words, we could consider non deterministic codewords, but those which take some values $x_1 \cdots x_k$ with certain probabilities. In this sense, a codeword $w_i$ will be defined by $\sum_j p_j^i \overline{x}_j^i$, where each $\overline{x}_j^i$ is a codeword as before In fact, it is very easy to see that our previous proof also works in this more general case if the probabilities $p_j^i$ are of the particular form $p_j^i = p_{j_1}^i \cdots p_{j_k}^i$. That is, if the letters of each codeword are chosen independently. Therefore, if we denote $C_c^1(\mathcal{N})$ the classical capacity of the channel $\mathcal{N}$ when the encoder $\mathcal{A}$ is restricted to the use of random codewords of the form $\sum_j p_j^i \overline{x}_j^i$, with $p_j^i = p_{j_1}^i \cdots p_{j_k}^i$, our previous argument shows that

$$C_c^1(\mathcal{N}) = \max_{P=(p(x))_x} H(X : Y). \tag{3.3}$$

In order to consider the general classical capacity $C_c(\mathcal{N})$, we must allow Alice to consider all random codewords $\sum_j p_j^i \overline{x}_j^i$. One can see that we can write the classical capacity $C_c(\mathcal{N})$ as a regularization of the $C_c^1$ capacity:

THEOREM 3.2. *Given a classical channel $\mathcal{N} : \ell_1^n \to \ell_1^n$, one has*

$$C_c(\mathcal{N}) = \sup_k \frac{C_c^1(\otimes^k \mathcal{N})}{k}. \tag{3.4}$$

PROOF. Let us first assume that $C_c(\mathcal{N}) = c$. According to our definition of capacity in Section 4, for every $\epsilon > 0$ we can find two natural numbers $m, k_0 \in \mathbb{N}$ such that $\frac{m}{k_0} = c$, an encoder $\mathcal{A} : \ell_1^{2^m} \to \otimes^{k_0} \ell_1^n$ and a decoder $\mathcal{B} : \otimes^{k_0} \ell_1^n \to \ell_1^{2^m}$ so that $\|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_{k_0}} \circ \mathcal{A}\| < \epsilon$. But this picture can be understood as that in which we send $m$ bits of classical information by using the channel $\otimes^{k_0}\mathcal{N}$ only once in our protocol (so we are computing $C^1$ !). In particular, $C_c^1(\otimes^{k_0}\mathcal{N}) \geq m$. We immediately conclude that $\sup_k \frac{C_c^1(\otimes^k \mathcal{N})}{k} \geq \frac{C_c^1(\otimes^{k_0} \mathcal{N})}{k_0} \geq \frac{m}{k_0} = c$. Thus, $\sup_k \frac{C_c^1(\otimes^k \mathcal{N})}{k} \geq C_c(\mathcal{N})$. Conversely, let us assume that $\frac{C_c^1(\otimes^k \mathcal{N})}{k} = c$ for some $k$. This means that for every $\epsilon > 0$ there exist $m, k_1 \in \mathbb{N}$ such that $\frac{m}{k_1} = kc$, an encoder $\mathcal{A} : \ell_1^{2^m} \to \otimes^{k_1}(\otimes^k \ell_1^n)$ and a decoder $\mathcal{B} : \otimes^{k_1}(\otimes^k \ell_1^n) \to \ell_1^{2^m}$ so that $\|id_{\ell_1^{2^m}} - \mathcal{B} \circ (\otimes^k \mathcal{N})^{\otimes_{k_1}} \circ \mathcal{A}\| < \epsilon$. This can be seen as a general protocol in which we are sending $m$ bits of classical communication by using $k_1 k$ times our channel $\mathcal{N}$. Note, however, that in order to understand the picture in this way, we must allow for general encoders $\mathcal{A} : \ell_1^{2^m} \to \ell_1^{n^{k_1 k}}$. This means that $C_c(\mathcal{N}) \geq \frac{m}{k_1 k} = c$ and, so, Theorem 3.2 is proved. □

The key point for us is that

$$C_c^1(\otimes^k \mathcal{N}) = k C_c^1(\mathcal{N}), \tag{3.5}$$

from where we deduce, via the previous theorem, that $C_c(\mathcal{N}) = C_c^1(\mathcal{N})$ and Theorem 3.1 follows from Equation (3.3).

According to (3.3), in order to prove (3.5) it suffices to show that

$$\max_{P=(p(x_1,x_2))_{x_1,x_2}} H(X_1, X_2 : Y_1, Y_2) = \max_{(p(x_1))_{x_1}} H(X_1 : Y_1) + \max_{(p(x_2))_{x_2}} H(X_2 : Y_2), \tag{3.6}$$

where $(X_1, X_2)$ denotes a random variable at the input of a channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ and $Y_1, Y_2$ is the induced random variable at the outputs of the channel (when $\mathcal{N}_1$ is a channel acting on the first variable $X_1$ and $\mathcal{N}_2$ another channel acting on the second variable $X_2$). An inductive argument allows us to obtain (3.5) from (3.6).

Inequality $\geq$ is in fact very easy. If $(p^*(x_1))_{x_1}$ and $(p^*(x_2))_{x_2})$ are the corresponding elements optimizing the right hand side term in (3.6), it is straightforward to check that $H(X_1, X_2 : Y_1, Y_2)$ yields to the same quantity when we consider the particular element $(p(x_1, x_2))_{x_1,x_2} = (p^*(x_1)p^*(x_2))_{x_1,x_2}$.

Let us now assume that $(p^*(x_1, x_2))_{x_1,x_2}$ is the element optimizing the left hand side term in (3.6). We will show that the probability distribution $(p^*(x_1)p^*(x_2))_{x_1,x_2}$ gives a value $H(X_1, X_2 : Y_1, Y_2)$ as good as $(p^*(x_1, x_2))_{x_1,x_2}$ where $(p^*(x_1))_{x_1}$ and $(p^*(x_2))_{x_2}$ denote the marginal distributions.

Recall that the action of $\mathcal{N}_1 \otimes \mathcal{N}_2$ is given by

$$p(y_1, y_2|x_1, x_2) = p(y_1|x_1)p(y_2|x_2).$$

If we sum over $y_2$ in the previous identity we obtain

$$p(y_1|x_1, x_2) = p(y_1|x_1).$$

That is, $Y_1$ conditioned to $X_1$ is independent of $X_2$. In the same way we can see that $Y_2$ conditioned to $X_2$ is independent of $X_1$ and $Y_1$. Indeed,

$$\begin{aligned}
p(y_2|x_1, x_2, y_1) &= \frac{p(x_1, x_2, y_1, y_2)}{p(x_1, x_2, y_1)} \\
&= \frac{p(y_1, y_2|x_1, x_2)p(x_1, x_2)}{p(y_1|x_1, x_2)p(x_1, x_2)} \\
&= \frac{p(y_1|x_1)p(y_2|x_2)}{p(y_1|x_1)} = p(y_2|x_2).
\end{aligned}$$

Then, according to the chaining rule and the subadditivity of the Shannon entropy

$$\begin{aligned}
H(X_1, X_2 : Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2, Y_1) \\
&= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&= H(X_1 : Y_1) + H(X_2 : Y_2).
\end{aligned}$$

CHAPTER 7

# Quantum Shannon Theory

## 1. Von Neumann entropy

In order to generalize the previous results to the quantum information context, we will imagine a source that prepares messages with $n$ letters, but where each letter is chosen from an ensemble of quantum states. The signal alphabet consists of a set of quantum states $\rho_x$, each occurring with a probability $p_x$. Then, the system is completely characterized by the density matrix

$$\rho = \sum_x p_x \rho_x.$$

For a density matrix $\rho$ we may define the *von Neumann entropy*

$$S(\rho) = -tr(\rho \log \rho).$$

Note that if we choose an orthonormal basis $\{|a\rangle\}$ that diagonalizes $\rho$, $\rho = \sum_a \lambda_a |a\rangle\langle a|$, we have

$$S(\rho) = H(A),$$

where $H(A)$ is the Shannon entropy of the ensemble $A = \{a, \lambda_a\}$.

We will see that von Neumann entropy plays a dual role in quantum information. On the one hand, it quantifies the (quantum) incompressible information content per letter of the quantum source or ensemble (in the case where the signal states are pure) in the same way as the Shannon entropy quantifies the information content of a classical source. That is, it quantifies the minimum number of qubits per letter needed to reliable encode the information. However, the von Neumann entropy also quantifies the classical information content (the maximal amount of information per letter, in bits, that we can gain about the preparation by making the best possible measurement).

We start this study by showing some important properties of the von Neumann entropy.

1. *Purity*: $S(\rho) = 0$ for every pure state $\rho = |\psi\rangle\langle\psi|$.
   This is trivial by definition.

2. *Invariance*: $S(U\rho U^*) = S(\rho)$ for every state and every unitary $U$.
   It is obvious since $S(\rho)$ depends only on the eigenvalues of $\rho$.

3. *Maximum*: $0 \leq S(\rho) \leq \log d$ for every state $\rho$ in dimension $d$, with equality when all the eigenvalues are equal.
   The proof follows from Equation (1.1).

4. For every pair of states $\rho$ and $\sigma$ we have

   $$S(\rho \otimes \sigma) = S(\rho) + S(\sigma).$$

   The proof follows form the fact that if $(\lambda_i)_i$ and $(\beta_j)_j$ are the eigenvalues of $\rho$ and $\sigma$ respectively, then $(\lambda_i \beta_j)_{i,j}$ are the eigenvalues of $\rho \otimes \sigma$.

5. Let $(p_i)_{i=1}^n$ be a probability distribution and $(\rho_i)_{i=1}^n$ be a family of states with mutually orthogonal supports. Then,

$$S\big(\sum_{i=1}^n p_i\rho_i\big) = H((p_i)_i) + \sum_{i=1}^n p_i S(\rho_i).$$

In particular,

$$S\big(\sum_{i=1}^n p_i|i\rangle\langle i|\otimes\rho_i\big) = H((p_i)_i) + \sum_{i=1}^n p_i S(\rho_i)$$

for every probability distribution $(p_i)_{i=1}^n$ and every family of states $(\rho_i)_{i=1}^n$.

Proof. Let us write $\rho_i = \sum_j \lambda_j^i |e_j^i\rangle\langle e_j^i|$, where $(\lambda_j^i)_j$ and $(|e_j^i\rangle)_j$ are the corresponding eigenvalues and eigenvectors of $\rho_i$ for every $i$. Then, because of our orthogonality condition we know that $(p_i\lambda_i^j)_{i,j}$ and $(|e_j^i\rangle)_{i,j}$ are the corresponding eigenvalues and eigenvectors of $\sum_{i=1}^n p_i\rho_i$. Therefore,

$$\begin{aligned}S\big(\sum_{i=1}^n p_i\rho_i\big) &= -\sum_{i,j} p_i\lambda_i^j \log(p_i\lambda_i^j) = -\sum_{i,j} p_i\lambda_i^j\log(p_i) - \sum_{i,j} p_i\lambda_i^j\log(\lambda_i^j)\\ &= -\sum_i p_i\log(p_i) - \sum_i p_i\sum_j \lambda_i^j\log(\lambda_i^j) = H((p_i)_i) + \sum_{i=1}^n p_iS(\rho_i).\end{aligned}$$

The second part of the statement follows from the fact that the states $|i\rangle\langle i|\otimes\rho_i$'s have mutually orthogonal supports and $S(|i\rangle\langle i|\otimes\rho_i) = S(\rho_i)$ for every $i$. □

In general, we have

$$S\big(\sum_{i=1}^n p_i\rho_i\big) \leq H((p_i)_i) + \sum_{i=1}^n p_iS(\rho_i)$$

for every family of states $(\rho_i)_{i=1}^n$. We refer [**8**, Theorem 11.10] for the proof.

6. *Klein's inequality*: Given two states $\rho$ and $\sigma$ we have that

$$S(\rho\|\sigma) := tr(\rho\log\rho) - tr(\rho\log\sigma) \geq 0,$$

with equality if and only if $\rho=\sigma$. The quantity $S(\rho\|\sigma)$ is known as the *relative entropy of $\rho$ to $\sigma$*.

Proof. Let $\rho = \sum_i p_i|e_i\rangle\langle e_i|$ and $\sigma = \sum_j q_j|f_j\rangle\langle f_j|$ with $(p_i)_i$ and $(q_j)_j$ probability distributions and $(|e_i\rangle)_i$ and $(|f_j\rangle)_j$ orthogonal basis. Then,

$$\begin{aligned}S(\rho\|\sigma) &= \sum_i p_i\log p_i - \sum_i\langle e_i|\rho\log\sigma|e_i\rangle = \sum_i p_i\log p_i - \sum_i p_i\langle e_i|\log\sigma|e_i\rangle\\ &= \sum_i p_i\log p_i - \sum_i p_i\langle e_i|\Big(\sum_j \log q_j|f_j\rangle\langle f_j|\Big)|e_i\rangle\\ &= \sum_i p_i\Big[\log p_i - \sum_j C_{i,j}\log q_j\Big],\end{aligned}$$

where $C_{i,j} = \langle e_i|f_j\rangle\langle f_j|e_i\rangle = |\langle e_i, f_j\rangle|^2$.

$(C_{i,j})_{i,j}$ is a double stochastic matrix. That is, $C_{i,j}\geq 0$, $\sum_i C_{i,j} = 1$ for every $j$ and $\sum_j C_{i,j} = 1$ for every $i$. Then, on the one hand we can define the probability distribution $(r_i)_{i=1}^n$ such that $r_i = \sum_j C_{i,j} q_j$ for every $i$ and, on the other hand, we can use the concavity property of the function $f(x) = \log(x)$ to state

$$\sum_j C_{i,j} \log q_j \leq \log r_i$$

for every $i$; with equality if for every $i$ there exists a value $j$ such that $C_{i,j} = 1$. That is, $(C_{i,j})_{i,j}$ is a permutation matrix. Therefore, we obtain that

$$S(\rho\|\sigma) \geq \sum_i p_i[\log p_i - \log r_i] = -\sum_i p_i \log \frac{r_i}{p_i}.$$

In order to prove that this quantity is not negative, we invoke inequality (1.2) to state

$$-\sum_i p_i \log \frac{r_i}{p_i} \geq \frac{1}{\ln 2} \sum_i p_i (1 - \frac{r_i}{p_i}) = \frac{1}{\ln 2} \sum_i (p_i - r_i) = 0,$$

with equality if and only if $p_i = r_i$ for every $i$.

The result follows now easily. □

7. *Subadditivity*: If we have a bipartite quantum state $\rho_{AB}$, then

$$S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B),$$

where $\rho_A = tr_B(\rho)$ and $\rho_B = tr_A(\rho)$. Moreover, the previous inequality is an equality (just) for uncorrelated systems $\rho = \rho_A \otimes \rho_B$.

PROOF. The proof of this result is a direct application of Klein's inequality. Indeed, if we denote $\rho = \rho_{AB}$ and $\sigma = \rho_A \otimes \rho_B$ we just note that

$$tr(\rho \log \rho) = -S(\rho_{AB}),$$

and

$$-tr(\rho \log \sigma) = -tr(\rho_{AB}(\log \rho_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes \log \rho_B)) = S(\rho_A) + S(\rho_B).$$

□

In fact, one can prove a much stronger result known as *Strong Subadditivity*:

THEOREM 1.1 (Strong Subadditivity). *If we have a tripartite quantum state $\rho_{ABC}$, then*

$$S(\rho_{ABC}) + S(\rho_B) \leq S(\rho_{AB}) + S(\rho_{BC}).$$

However, the proof of this result is much more difficult than the previous one. We refer [**8**, Section 11.4] for its proof.

8. *Concavity*: For every probability distribution $(p_i)_{i=1}^n$ and $(\rho_i)_{i=1}^n$ family of states we have

$$S(\sum_{i=1}^n p_i \rho_i) \geq \sum_{i=1}^n p_i S(\rho_i).$$

PROOF. Concavity can be easily obtained from the subadditivity property shown above. Indeed, let us define the state $\rho_{AB} = \sum_i p_i \rho_i \otimes |i\rangle\langle i|$. According to Property 5. we have

$$H((p_i)_i) + \sum_i p_i S(\rho_i) = S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B) = S(\sum_i p_i \rho_i) + H((p_i)_i).$$

□

9. *Triangle inequality (Araki-Lieb inequality)*: If we have a bipartite quantum state $\rho_{AB}$, then

$$S(\rho_{AB}) \geq \big|S(\rho_A) - S(\rho_B)\big|.$$

Proof. Let us consider a system $R$ which purifies the systems $A$ and $B$. Then, according to the subadditivity of the von Neumann entropy, we have $S(\rho_{AR}) \leq S(\rho_A) + S(\rho_R)$. Now, according to Section 3 in Chapter 4, we know that $S(\rho_{AR}) = S(\rho_B)$ and that $S(\rho_R) = S(\rho_{AB})$. Then, we conclude that $S(\rho_{AB}) \geq S(\rho_B) - S(\rho_A)$. □

There is a remarkable point here. The analogous property for the Shannon entropy is $H(X,Y) \geq \max\{H(X), H(Y)\}$ (or, equivalently, $H(X|Y), H(X|Y) \geq 0$). Note that this is not true in the quantum setting. Indeed, for a bipartite entangled pure state $\rho_{AB}$, where $S(\rho_A) = S(\rho_B) \neq 0$, we have $0 = S(\rho_{AB}) < S(\rho_A)$. We see that this state $\rho_{AB}$ has a definite preparation, but if we measure observables of the subsystems, the measurement outcomes are inevitably random and unpredictable. We cannot discern how the state was prepared by observing the two subsystems separately, rather, information is encoded in the nonlocal quantum correlation.

## 2. Schumacher's compression theorem

In order to study how to compress quantum information, we must start by explaining what we understand by a quantum source. Analogously to the classical case, we will consider long messages consisting of $n$ letters. However, in this case each letter will be chosen at random from the ensemble of pure states

$$\{|\varphi_x\rangle\langle\varphi_x|, p_x\},$$

where the $|\varphi_x\rangle$'s are not necessarily mutually orthogonal. Thus, each letter is described by the density matrix

$$\rho = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|,$$

and the entire message has the density matrix

$$\rho^n = \rho \otimes \cdots \otimes \rho.$$

Note that a quantum source is not, in principle, characterized by a quantum state $\rho$, since this state could be written in two different ways:

$$\rho = \sum_x p_x |\varphi_x\rangle\langle\varphi_x| = \sum_y q_y |\psi_y\rangle\langle\psi_y|.$$

In fact, one of these ways is always its spectral decomposition $\rho = \sum_i \lambda_i |i\rangle\langle i|$. However, Schumacher's compression theorem will tell us that in order to quantify the quantum incompressible information content per letter of the quantum source the only relevant information is the quantum state $\rho$ or, more precisely, its von Neumann entropy $S(\rho)$. In particular, all quantum sources defined by the same quantum state will give rise to the same quantum incompressible information.

Before going on, one could wonder in which sense our definition of a quantum source extends the idea of a classical source. That is, how can one "perform" a classical source of information $X = \{a, p_a\}$ by using a quantum source? In order to do this, we will need to extract classical information from a quantum state and this is usually done by means of a measurement. Let us consider the state

$$\rho = \sum_a p_a |a\rangle\langle a|,$$

where $\rho$ is a density matrix on a Hilbert space defined by the orthonormal basis $\{|a\rangle\}_a$. In particular, we can understand $\rho$ as the state associated to the quantum source $\{|a\rangle\langle a|, p_a\}$.

Now, we define the (von Neumann) measurement $\{E_a = |a\rangle\langle a|\}_a$. Then, we can extract a letter from our classical alphabet by measuring on $\rho$ with our measurement $\{E_a\}_a$. Note that we have

$$p(a) = tr(E_a\rho) = p_a \quad \text{for every} \quad a.$$

Moreover, we have that $S(\rho) = H(X)$. That is, we can describe our classical source by means of a measurement process associated to a quantum source. Conversely, in the case where the signal alphabet of a quantum source consists of mutually orthogonal pure states, the quantum source reduces to a classical one, since all of the signal states can be perfectly distinguished and $S(\rho) = H(X)$.

The process of extracting classical information from a quantum state by means of a measurement is very important in quantum information theory. In general, we will imagine a source that prepares messages with $n$ letters, but where each letter is chosen from an ensemble of quantum states. The signal alphabet consists of a set of quantum states $\rho_x$[1], each occurring with a probability $p_x$. Then, the probability of any outcome of any measurement of a letter chosen from this ensemble, if the observer has no knowledge about which letter was prepared, can be complete characterized by the density matrix $\rho = \sum_x p_x\rho_x$; for the POVM $\{F_a\}_a$ we have $Prob(a) = tr(F_a\rho)$. This scenario will be very important in the following sections.

A *compression scheme of rate $R$* for this source consists of two families of quantum operations (quantum channels) $C^n$ and $D^n$. $C^n$ is the compression operation, taking states in $H^{\otimes n}$ to states in a $2^{nR}$-dimensional space, the *compressed space*. We may regard the compressed space as representing $nR$ qubits. $D^n$ is the decompression operation, which takes states in the compressed space to states in the the original space. Finally, we must also define a good criterium of reliability between (all) our messages $\rho^n$ and its image under the compression and decompression operations $D^n \circ C^n(\rho^n)$.

A standard measure of how close two density operators $\rho$ and $\sigma$ are is given by the *fidelity*:

$$F(\rho,\sigma) = tr\sqrt{\rho^{\frac{1}{2}}\sigma\rho^{\frac{1}{2}}}.$$

Note that if, say, $\sigma = |\varphi\rangle\langle\varphi|$ is a pure state, we have

$$F(\rho,\sigma) = \sqrt{\langle\varphi|\rho|\varphi\rangle}.$$

In fact, this nice measure is known to be "equivalent" to the trace distance:

$$1 - F(\rho,\sigma) \leq D(\rho,\sigma) \leq \sqrt{1 - F(\rho,\sigma)^2},$$

with $\frac{1}{2}\|\rho - \sigma\|_1$.

Therefore, we could impose as our criterium of reliability that $F(\rho, D^n \circ C^n(\rho))$ is very close to one for a large $n$. However, this distance doesn't really capture the idea of preserving the quantum information from a quantum source through the action of a quantum channel[2]. Note that although we are "encoding" all the messages by using just one state $\rho^n$, in order to preserve the information from the corresponding quantum source we must read this operator as

$$\rho^n = \sum_{x_1,\cdots,x_n} p_{x_1}\cdots p_{x_n}|\varphi_{x_1}\rangle\langle\varphi_{x_1}| \otimes \cdots \otimes |\varphi_{x_n}\rangle\langle\varphi_{x_n}|,$$

and we must preserve the quantum information "given by each of the terms"

$$p_{x_1}\cdots p_{x_n}|\varphi_{x_1}\rangle\langle\varphi_{x_1}| \otimes \cdots \otimes |\varphi_{x_n}\rangle\langle\varphi_{x_n}|.$$

This extends the idea of compressing a classical source.

[1] Here, we don't have to restrict to the case of pure states.

[2] In fact, if we choose the standard fidelity as our criterium of reliability in Schumacher's compression theorem, we could give a trivial proof by using a rank-one encoder and decoder.

According to our definition of a quantum source, a more suitable measure of reliability for a quantum source $\rho = \sum_i p_i \rho_i$ under a quantum channel $\mathcal{N}$ is the *ensemble average fidelity*

$$\overline{F} = \sum_j p_j F(\rho_j, \mathcal{N}(\rho_j))^2.$$

In particular, this notion extends our criterium of reliability in the classical context (Theorem 1.2).

However, there exits another "better" measure that we will use here, namely the *the entanglement fidelity.* This measure imposes that the entanglement of our state with the environment must be well preserved under the action of the channel $\mathcal{N}$. Therefore, this measure emphasizes the understanding of quantum entanglement as a measure of information. To define the entanglement fidelity, given a state $\rho$ acting on a Hilbert space $H_R$, we consider a purification $|\varphi\rangle \in H_R \otimes H_Q$ of $\rho$ as we explained in Chapter 4, Section 3. Then, we define the entanglement fidelity of $\rho$ and the channel $\mathcal{N}$ by

$$F(\rho, \mathcal{N}) = F(QR, QR') = \langle\varphi|(\mathcal{N} \otimes \mathbb{1}_R)(|\varphi\rangle\langle\varphi|)|\varphi\rangle^3.$$

A very important fact for us is that if our channel $\mathcal{N}$ is given by a set $\{E_i\}_i$ of Krauss operators (see Theorem 4.1), one can show that

$$F(\rho, \mathcal{N}) = \sum_i |tr(\rho E_i)|^2. \tag{2.1}$$

Proof.

$$F(\rho, \mathcal{N}) = \langle QR|\rho^{Q'R'}|QR\rangle = \sum_i |\langle QR|E_i \otimes \mathbb{1}|QR\rangle|^2.$$

Let us consider $|QR\rangle = \sum_j \sqrt{p_j j}\rangle|j\rangle$, where $\rho = \sum_j p_j|j\rangle\langle j|$. Then,

$$\begin{aligned}\langle QR|E_i \otimes \mathbb{1}|QR\rangle &= \sum_{j,k} \sqrt{p_j}\sqrt{p_k}\langle j|E_i|k\rangle\langle j|k\rangle \\ &= \sum_j p_j\langle j|E_i|j\rangle = tr(E_i\rho).\end{aligned}$$

Therefore,

$$F(\rho, \mathcal{N}) = \sum_i |\langle QR|E_i \otimes \mathbb{1}|QR\rangle|^2 = \sum_i |tr(E_i\rho)|^2.$$

□

The fact that we use the expression "better" when we compare the entanglement fidelity with the average fidelity is because it can be proved that

$$F(\sum_j p_j\rho_j, \mathcal{N}) \leq \overline{F}. \tag{2.2}$$

The proof of (2.2) makes use of the fact that

$$F(\rho, \mathcal{N}) \leq F(\rho, \mathcal{N}(\rho))^2 \tag{2.3}$$

which is very easy to see[4].

[4] Intuitively, this means that it is harder to preserve a state plus its entanglement with the environment than just preserving the state.

Proof. We will prove that, for a fixed channel $\mathcal{N}$ the entanglement fidelity $F(\rho,\mathcal{N})$ is a convex function of $\rho$. Then, according to (2.3), it will follow that for every $\rho=\sum_j p_j\rho_j$ we have

$$F(\rho,\mathcal{N})\leq\sum_j p_jF(\rho_j,\mathcal{N})\leq\sum_j p_jF(\rho_j,\mathcal{N}(\rho_j))^2=\overline{F}.$$

Let us consider two states $\rho_1$ and $\rho_2$ and let us define the function

$$F(\lambda)=F(\lambda\rho_1+(1-\lambda)\rho_2,\mathcal{N})=\sum_i\left|tr\big((\lambda\rho_1+(1-\lambda)\rho_2)E_i\big)\right|^2,$$

where we have used Equation (2.1) in the last equality. Then, by expanding each of therm in the previous sum $\left|tr\big((\lambda\rho_1+(1-\lambda)\rho_2)E_i\big)\right|^2$ as

$$\lambda^2\left|tr\big(\rho_1E_i\big)\right|^2+\lambda(1-\lambda)\Big(tr\big(\rho_1E_i\big)\overline{tr\big(\rho_2E_i\big)}+tr\big(\rho_2E_i\big)\overline{tr\big(\rho_1E_i\big)}\Big)+(1-\lambda)^2\left|tr\big(\rho_2E_i\big)\right|^2,$$

we easily obtain that

$$F''(\lambda)=2\sum_i\left|tr\big((\rho_1-\rho_2)E_i\big)\right|^2\geq 0.$$

Therefore, $F(\cdot)$ is a convex function of $\lambda$. We deduce the convexity of the entanglement fidelity $F(\rho,\mathcal{N})$ from here since

$$\begin{aligned}F(\lambda\rho_1+(1-\lambda)\rho_2,\mathcal{N})=F(\lambda1+(1-\lambda)0)&\leq\lambda F(1)+(1-\lambda)F(0)\\&=\lambda F(\rho_1,\mathcal{N})+(1-\lambda)F(\rho_2,\mathcal{N}).\end{aligned}$$

This concludes the proof. □

Therefore, having a high fidelity for the entanglement fidelity implies a high fidelity for the average fidelity. This point, joint with the nice expression (2.1), are the main reasons to choose the entanglement fidelity as our criterium of reliability in the Schumacher compression theorem.

The basic idea to prove Schumacher's compression theorem is to extend the idea of typical sequences to that of typical subspaces. Let us decompose our state as

$$\rho=\sum_x p_x|x\rangle\langle x|,$$

where $(|x\rangle)_x$ is an orthonormal set and $(p(x))_x$ are the eigenvalues of $\rho$. Note that $(p(x))_x$ is a probability distribution such that $H((p(x))_x)=S(\rho)$. Therefore, it makes sense to talk of an $\epsilon$-typical sequence exactly in the same way as in the classical case. In fact, we may say that a state $|x_1\rangle|x_2\rangle\cdots|x_n\rangle$ is $\epsilon$-typical id the sequence $x_1x_2\cdots x_n$ is $\epsilon$-typical. Furthermore, we define the $\epsilon$-typical subspace as the subspace spanned by the $\epsilon$-typical states. We will denote this subspace by $T(n,\epsilon)$, and the projector onto the $\epsilon$-typical subspace by $P(n,\epsilon)$. Notice that

$$P(n,\epsilon)=\sum_{x\ \epsilon\text{-typical}}|x_1\rangle\langle x_1|\otimes|x_2\rangle\langle x_2|\otimes\cdots\otimes|x_n\rangle\langle x_n|.$$

The following theorem is a simple consequence of the Theorem of typical sequences.

Theorem 2.1 (Typical subspace theorem).

1. *Fix $\epsilon>0$. Then, for any $\delta>0$, for sufficiently large $n$,*

$$tr\big(P(n,\epsilon)\rho^n\big)\geq 1-\delta.$$

2. *For any fixed $\epsilon>0$ and $\delta>0$, for sufficiently large $n$, the dimension $|T(n,\epsilon)|=tr\big(P(n,\epsilon)\big)$ of $T(n,\epsilon)$ satisfies*

$$(1-\delta)2^{n\left(S(\rho)-\epsilon\right)}\leq|T(n,\epsilon)|\leq 2^{n\left(S(\rho)+\epsilon\right)}.$$

3. *Let $S(n)$ be a projector onto any subspace of $H^{\otimes n}$ of dimension at most $2^{nR}$, where $R < S(\rho)$ is fixed. Then, for any $\delta > 0$ and sufficiently large $n$,*
$$tr\big(S(n)\rho^n\big) \leq \delta.$$

Proof. The first point of the theorem is a direct consequence of the first part of the Theorem of typical sequences and the fact that

The second point follows trivially from the second part of the Theorem of typical sequences and the fact that $|T(n,\epsilon)|$ is the same number in both cases.

$$tr\big(P(n,\epsilon)\rho^n\big) = \sum_{x\ \epsilon\text{-typical}} p(x_1)p(x_2)\cdots p(x_n).$$

Finally, in order to prove the third part we split the trace up as

$$tr\big(S(n)\rho^n P(n,\epsilon)\big) + tr\big(S(n)\rho^n(\mathbb{1} - P(n,\epsilon))\big)$$

To upper bound the first term, we first note that, by definition, $P(n,\epsilon)$ is a projector which commutes with $\rho^n$. Then,

$$tr\big(S(n)\rho^n P(n,\epsilon)\big) = tr\big(S(n)P(n,\epsilon)\rho^n P(n,\epsilon)\big) \leq 2^{nR}2^{-n\left(S(\rho)-\epsilon\right)},$$

since the eigenvalues of $P(n,\epsilon)\rho^n P(n,\epsilon)$ are bounded above by $2^{-n\left(S(\rho)-\epsilon\right)}$. Letting $n \to \infty$ we see that the this term goes to zero. On the other hand, using that $S(n) \leq \mathbb{1}$ and that $S(n)$ and $\rho^n(I - P(n,\epsilon))$ are both positive operators, it follows that

$$0 \leq tr\big(S(n)\rho^n(\mathbb{1} - P(n,\epsilon))\big) \leq tr\big(\rho^n(\mathbb{1} - P(n,\epsilon))\big),$$

which tends to zero as $n \to \infty$. The result follows now trivially. □

Theorem 2.2 (Schumacher's noiseless channel coding theorem). *Let $\{H,\rho\}$ be an i.i.d. quantum source. If $R > S(\rho)$, then there exists a reliable compression scheme of rate $R$ for the source. Conversely, if $R < S(\rho)$, then any compression scheme will not be reliable.*

Proof. Suppose $R > S(\rho)$ and let $\epsilon > 0$ be such that $S(\rho) + \epsilon \leq R$. By the typical subspace theorem, for any $\delta > 0$ and for sufficiently large $n$, we have $tr\big(P(n,\epsilon)\rho^n\big) \geq 1 - \delta$, and $\dim\big(T(n,\epsilon)\big) = tr\big(P(n,\epsilon)\big) \leq 2^{nR}$. Let $H_c^n$ be a Hilbert space of dimension $2^{nR}$ containing the subspace $T(n,\epsilon)$. The encoding will be defined in the following way. First, we apply the projective measurement defined $P(n,\epsilon)$, $\mathbb{1} - P(n,\epsilon)$, with corresponding outputs 0 and 1 respectively, on the corresponding state. If the output 0 occurs, then we don't do anything and the state is left in the subspace $T(n,\epsilon)$. On the other hand, if we obtain the output 1, then we just replace the state of the system with some standard state $|0\rangle$ chosen from the typical subspace. It doesn't matter which state is used. It follows that the encoding map $C^n$, which will send states acting on $H^{\otimes n}$ into states acting on $H_c^n$, is defined by

$$C^n(\sigma) = P(n,\epsilon)\sigma P(n,\epsilon) + \sum_i A_i\sigma A_i^\dagger,$$

where, $A_i = |0\rangle\langle i|$ and $|i\rangle$ is an orthonormal basis for the orthocomplement of the typical subspace. On the other hand, the decoding operation $D^n$, which will send states action on $H_c^n$ into states acting on $H^{\otimes n}$, is defined simply as the identity (inclusion) operator $D^n(\sigma) = \sigma$[5]. In order to lower bound the entangled fidelity we have

$$F(\rho^n, D^n \circ C^n) = |tr\big(\rho^n P(n,\epsilon)\big)|^2 + \sum_i |tr\big(\rho^n A_i\big)|^2 \geq |tr\big(\rho^n P(n,\epsilon)\big)|^2 \geq (1-\delta)^2 \geq 1 - 2\delta.$$

Since this is for an arbitrary $\delta$ we conclude the first part of the theorem.

[5] Note that the definition of our encoder and our decoder makes use of the realizations of $B(H_c^n)$ and $B(T(n,\epsilon))$ as certain subalgebras of $B(H^{\otimes n})$ and the corresponding identification of the states.

Let us now assume that $R < S(\rho)$. Let us also denote $\{C^n, D^n\}$ the corresponding encoder-decoder of a compression scheme. With out loss of generality we can assume that the compression operator $C^n$ maps states acting on $H^{\otimes n}$ to states acting on a subspace $H_0^n \subset H^{\otimes n}$ of dimension $2^{nR}$. Let us also denote $S(n) : H^{\otimes n} \to H_0^n \subset H^{\otimes n}$ the corresponding projector. Finally, let us denote by $\{C_j\}_j$ and $\{D_k\}_k$ some operation elements for the channels $C^n$ and $D^n$ respectively. Then, we can write

$$F(\rho^n, D^n \circ C^n) = \sum_{j,k} |tr(D_k C_j \rho^n)|^2.$$

Note that $C_j : H^{\otimes n} \to H_0^n$ for every $j$ and $D_k : H_0^n \to H^{\otimes n}$ for every $k$. On the other hand, if we denote by $S^k(n) : H^{\otimes n} \to D_k(H_0^n) \subset H^{\otimes n}$ the projector onto the subspace $D_k(H_0^n)$, we clearly have the identity $D_k C_j = S^k(n) D_k C_j$. Thus,

$$\begin{aligned} F(\rho^n, D^n \circ C^n) = \sum_{j,k} |tr(D_k C_j \rho^n)|^2 &= \sum_{j,k} |tr(D_k C_j \rho^n S^k(n))|^2 \\ &\leq \sum_{j,k} tr(D_k C_j \rho^n C_j^\dagger D_k^\dagger) tr(\rho^n S^k(n)), \end{aligned}$$

where we have applied Cauchy-Schwartz inequality in the last step. Now, by the third point of the typical subspace theorem, we know that for any $\delta$ and sufficiently large $n$, we must have $|tr(\rho^n S^k(n))| \leq \delta$ independently of $k$. Therefore,

$$F(\rho^n, D^n \circ C^n) \leq \delta \sum_{j,k} tr(D_k C_j \rho^n C_j^\dagger D_k^\dagger) = \delta,$$

since $C^n$ and $D^n$ are trace preserving maps. Again, since $\delta$ is arbitrary, it follows that $F(\rho^n, D^n \circ C^n) \to 0$ as $n \to \infty$; and thus compression is not reliable. □

Note that the previous proof does not say anything about the second part $R < S(\rho)$ for the average fidelity. Indeed, Equation (2.2) allows us to conclude that $\overline{F}(\rho^n, D^n \circ C^n) \geq 1 - 2\delta$, where

$$\rho^n = \sum_{x_1, \cdots, x_n} p_{x_1} \cdots p_{x_n} |\varphi_{x_1}\rangle\langle\varphi_{x_1}| \otimes \cdots \otimes |\varphi_{x_n}\rangle\langle\varphi_{x_n}|,$$

in the case $R > S(\rho)$. However, we cannot obtain any upper bound for $\overline{F}(\rho^n, D^n \circ C^n)$ just looking at $F(\rho^n, D^n \circ C^n)$. In the following we present an argument to show that compressing quantum information at a rate $R$ lower that $S(\rho)$ is not possible when we consider the average fidelity as our criterium of reliability. The argument is just a slight modification of the previous proof.

In order to simplify notation let us write $\rho^n = \sum_{\bar{x}} p_{\bar{x}} |\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|$ to denote the state above. As before, let us also denote $\{C^n, D^n\}$ the corresponding encoder-decoder of a compression scheme. With out loss of generality we can assume that the compression operator $C^n$ maps states acting on $H^{\otimes n}$ to states acting on a subspace $H_0^n \subset H^{\otimes n}$ of dimension $2^{nR}$. Let us also denote $S(n) : H^{\otimes n} \to H_0^n \subset H^{\otimes n}$ the corresponding projector. We can use the concavity of the function $f(x)=\sqrt{x}$ to state

$$\begin{aligned}\overline{F}(\rho^n, D^n\circ C^n) &= \sum_{\bar{x}} p_{\bar{x}} F\big(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|, D^n\circ C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big)\\ &= \sum_{\bar{x}} p_{\bar{x}}\sqrt{\langle\varphi_{\bar{x}}|D^n\circ C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)|\varphi_{\bar{x}}\rangle}\\ &\leq \sqrt{\sum_{\bar{x}} p_{\bar{x}}\langle\varphi_{\bar{x}}|D^n\circ C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)|\varphi_{\bar{x}}\rangle}\\ &= \sqrt{\sum_{\bar{x}} p_{\bar{x}} tr\big(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|D^n\circ C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big)}.\end{aligned}$$

Now, note that

$$\begin{aligned}\sum_{\bar{x}} p_{\bar{x}} tr\big(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|D^n\circ C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big) &= \sum_{\bar{x}} p_{\bar{x}} tr\big((D^n)^*(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big)\\ &\leq \sum_{\bar{x}} p_{\bar{x}}\big\|(D^n)^*(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big\|_{B(H_0^n)} tr\big(C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big)\\ &\leq \sum_{\bar{x}} p_{\bar{x}} tr\big(S(n)C^n(|\varphi_{\bar{x}}\rangle\langle\varphi_{\bar{x}}|)\big)\\ &= tr\big(S(n)C^n(\rho^n)\big)\leq\delta.\end{aligned}$$

Indeed, the first equality follows from the definition of the transpose map. On the other hand, we have already mentioned that for every quantum channel $\mathcal{N}: S_1(H_A)\to S_1(H_B)$, its transpose map $\mathcal{N}^*: B(H_B)\to B(H_A)$ is a completely positive and unital map. In particular, $\|\mathcal{N}^*(\rho)\|_{B(H_A)}\leq 1$ for every state $\rho$. We have also used that $C^n(\rho)=S(n)C^n(\rho)S(n)$ for every state $\rho$, which follows from the definition of the projection $S(n)$. Finally, the last inequality follows from the third point in the Typical subspace theorem.

Therefore, even when we consider the average fidelity as our criterium of reliability, the fidelity of every encoder-decoder of a compression scheme at rate $R<S(\rho)$ tends to zero. It is worth mentioning that this last upper bound is stronger than the one provided in the proof of the previous theorem.

**2.1. Holevo Information.** The Schumacher theorem characterizes the compressibility of an ensemble of pure states. But, what if the letters are drawn from an ensemble of mixed states? Then, it is easy to see that $S(\rho)$ is not the right answer, as is shown in the following example:

We could consider the trivial example of a quantum source defined by just one mixed element $\rho$ such that $S(\rho)\neq 0$ to which we associate probability $p_0=1$. That is, the quantum source is $\{p_0=1,\rho\}$. Then, the message is always $\rho\otimes\cdots\otimes\rho$ and it carries no information. Indeed, Bob can reconstruct the message perfectly without receiving anything from Alice. Therefore, the message can be compressed to zero qubits per letter. which is less than $S(\rho)$.

Given an ensemble of mixed states $\mathcal{E}=\{(p_x)_x,(\rho_x)_x\}$, we defined the *Holevo information of the ensemble* $\mathcal{E}$ as

$$\mathcal{X}(\mathcal{E})=S\big(\sum_x p_x\rho_x\big)-\sum_x p_x S(\rho_x).$$

Note that this quantity reduces to the von Neumann entropy if $\rho_x=|\varphi_x\rangle\langle\varphi_x|$ is a pure state for every $x$. On the other hand, in the case in which the states $\rho_x$'s are mutually orthogonal (so they are perfectly distinguishable), $\mathcal{X}(\mathcal{E})$ reduces to the Shannon capacity $H(X)$ (this can be seen by considering a purification of the states $\rho_x$ or using Property 5 on the von Neumann entropy above). Therefore, in both cases, the Holevo information is the optimal number of qubits per letter that can be attained if we are to compress the message while retaining good fidelity for large $n$.

Up to our knowledge, it is not known if $\mathcal{X}(\mathcal{E})$ is the right quantity to measure the compressibility for an ensemble of mixed states. In fact, it can be seen that it is a lower bound for the compressibility ratio in this general case. That is, high-fidelity compression to less than $\mathcal{X}(\mathcal{E})$ qubits per letter is not possible. However, it is not known whether compression to $\mathcal{X}(\mathcal{E})$ qubits per letter is asymptotically attainable.

## 3. Accessible Information and Holevo bound

The previous section was devoted to quantifying the quantum information content -measured in qubits- of messages constructed from an alphabet of quantum states. In the following, we want to quantify the classical information -measured in bits- that *can be extracted from a source.*

Suppose Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{(p_x)_x, (\rho_x)_x\}$. Bob knows the ensemble but he doesn't know the particular state that Alice chose. The goal of Bob is to acquire as much information as possible about $x$. The way he collects the information is by performing a generalized measurement POVM $\{E_y\}_y$. Note, that if Alice chose the preparation $x$, Bob will obtain the measurement outcome $y$ with probability

$$p(y|x) = tr(E_y \rho_x).$$

This conditional probability, together with the ensemble $X$, determine the amount of information that Bob gains, on the average. As we explained in Section 2 this quantity is measured by the mutual information $H(X : Y)$ of the preparation and the measurement outcome. However, since Bob is free to perform any measurement, we are interested in his best possible choice. We define the *accessible information of the ensemble* $\mathcal{E} = \{(p_x)_x, (\rho_x)_x\}$ as

$$Acc(\mathcal{E}) = max_{\{E_y\}_y} H(X : Y).$$

It is important to note that in the case where the states $\rho_x$'s are perfectly distinguishable (so they are mutually orthogonal), we have an equality

$$Acc(\mathcal{E}) = H(X).$$

Note that, in particular, this happens in the classical case (this is an explanation of why the accessible information "makes no sense" in the classical context). In fact, this is an "if and only if" since one can easily check that one cannot have such an equality if the states of our ensemble are not mutually orthogonal. Actually, this fact can be understood as an equivalent formulation of the no-cloning theorem explained in Chapter 3 (see [**8**, page 530] for details).

The following theorem tells us that the Holevo information of an ensemble is an upper bound for the accessible information of such an ensemble.

THEOREM 3.1 (Holevo bound). *Suppose Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{(p_x)_x, (\rho_x)_x\}$. Then,*

$$Acc(\mathcal{E}) \leq \mathcal{X}(\mathcal{E}). \tag{3.1}$$

REMARK 3.2. *An immediate consequence of Holevo bound is that $Acc(\mathcal{E}) \leq \log n$ if we are dealing with states in dimension $n$. In other words, we cannot attain a classical capacity better than $n$ bits if we are dealing with $n$ qubits.*

PROOF. The proof of this theorem is based on an "artificial" construction involving three systems and the use of the strong subadditivity inequality. Let $X$ denote a quantum system with orthonormal basis $\{|x\rangle\}_x$, let $Q$ denote the quantum system that Alice gives to Bob and $Y$ an artificial quantum system with orthonormal basis $\{|y\rangle\}_y$. Let us consider the state

$$\rho = \sum_x p_x |x\rangle\langle x| \otimes \rho_x \otimes |0\rangle\langle 0|$$

in such a system. This state represents the situation in which Alice has chosen the state $\rho_x$ with probability $p_x$, she has given the state to Bob, who is about to use his measuring apparatus, initially in the standard state $|0\rangle$, to perform the measurement. By taking partial traces we see that

$$\rho_X = \sum_x |x\rangle\langle x| \quad \text{and} \quad \rho_Q = \sum_x \rho_x = \rho.$$

Therefore,

$$S(\rho_X) = H(X) \quad \text{and} \quad S(\rho_{QY}) = S(\rho_Q) = S(\rho).$$

Actually, since the $|x\rangle$'s are mutually orthogonal, we also have

$$S(\rho) = S(\rho_{XQ}) = \sum_x p_x S(\rho_x) + H(X).$$

Now, we will perform a unitary transformation that inputs Bob's measurements result in the output system $Y$. Let us first assume that Bob performs a von Neuamnn measurement: $\{P_y\}_y$, $P_y P_{y'} = \delta_{y,y'} P_y$ and we will treat the general case later.

Our unitary transformation is defined by

$$U_{QY}(|\sigma\rangle_Q \otimes |0\rangle_Y) = \sum_y P_y |\sigma\rangle_Q \otimes |y\rangle_Y.$$

In fact, it is very easy to see that this map preserves inner products and so, it can be extended to a unitary map on the space $QY$. Thus, this unitary acts on our state (with identity on the system $X$) as

$$\rho \to \rho' = \sum_{x,y,y'} p_x |x\rangle\langle x| \otimes P_y \rho_x P_{y'} \otimes |y\rangle\langle y'|.$$

Now, we invoke strong subadditivity in the form

$$S(\rho') + S(\rho'_Y) \leq S(\rho'_{XY}) + S(\rho'_{QY}). \tag{3.2}$$

In order to compute the corresponding entropies note that

$$S(\rho') = S(\rho) = \sum_x p_x S(\rho_x) + H(X) \quad \text{and} \quad S(\rho_{QY}) = S(\rho'_{QY}) = S(\rho).$$

Here, we have used that the von Neumann entropy is invariant under unitary transformations. On the other hand, since we are dealing with a von Neumann measurement, we deduce

$$\rho'_{XY} = \sum_{x,y} p_x tr(P_y \rho_x) |x\rangle\langle x| \otimes |y\rangle\langle y| = \sum_{x,y} p_x tr(P_y \rho_x) |xy\rangle\langle xy| = \sum_{x,y} p_{x,y} |xy\rangle\langle xy|.$$

Therefore,

$$S(\rho'_{XY}) = H(X, Y).$$

Now, we can also compute

$$\rho'_Y = \sum_y p(y) |y\rangle\langle y|, \quad \text{so} \quad S(\rho'_Y) = H(Y),$$

where we have used that $p(y) = \sum_x p(x) tr(P_y \rho_x)$.

Hence, Equation (3.2) becomes

$$\sum_x p_x S(\rho_x) + H(X) + H(Y) \leq H(X, Y) + S(\rho),$$

so

$$H(X:Y) = H(X) + H(Y) - H(X, Y) \leq S(\rho) - \sum_x p_x S(\rho_x) = \mathcal{X}(\mathcal{E}).$$

In order to conclude the proof, we must show how to obtain the result when Bob performs a general POVM $\{E_y\}_y$ rather than a von Neumann measurement. We can reduce this situation to the previous one by adding another subsystem $Z$ also initiated in a standard state $|0\rangle\langle 0|$:

$$\rho = \sum_x p_x |x\rangle\langle x| \otimes \rho_x \otimes |0\rangle\langle 0| \otimes |0\rangle\langle 0|.$$

Then, we see again that we can defined a unitary $U_{QYZ}$ such that

$$U_{QYZ}(|\sigma\rangle_Q \otimes |0\rangle_Y \otimes |0\rangle_Z) = \sum_y \sqrt{E_y}|\sigma\rangle_Q \otimes |y\rangle_Y \otimes |y\rangle_Z,$$

so that

$$\rho \to \rho' = \sum_{x,y,y'} p_x |x\rangle\langle x| \otimes \sqrt{E_y}\rho_x\sqrt{E_{y'}} \otimes |y\rangle\langle y'| \otimes |y\rangle\langle y'|.$$

Then, we invoke strong subadditivity in the form

$$S(\rho') + S(\rho'_Z) \leq S(\rho'_{XZ}) + S(\rho'_{QYZ}). \tag{3.3}$$

As before,

$$S(\rho') = S(\rho) = \sum_x p_x S(\rho_x) + H(X) \quad \text{and} \quad S(\rho'_{QYZ}) = S(\rho_{QYZ}) = S(\rho).$$

On the other hand, we have

$$S(\rho'_Z) = S\big(\sum_{x,y} p_x p(y|x)|y\rangle\langle y|\big) = S\big(\sum_{x,y} p_{x,y}|y\rangle\langle y|\big) = H(Y).$$

Finally,

$$\rho'_{XZ} = S\big(\sum_{x,y} p_x p(y|x)|x\rangle\langle x| \otimes |y\rangle\langle y| = S\big(\sum_{x,y} p_x p(y|x)|xy\rangle\langle xy| = H(X,Y).$$

Hence, Equation (3.3) becomes

$$\sum_x p_x S(\rho_x) + H(X) + H(Y) \leq H(X,Y) + S(\rho),$$

so

$$H(X:Y) = H(X) + H(Y) - H(X,Y) \leq S(\rho) - \sum_x p_x S(\rho_x) = \mathcal{X}(\mathcal{E}).$$

This concludes the proof. □

Remark 3.3.

1. *In the case of pure states $\rho_x = |\varphi_x\rangle\langle\varphi_x|$, the previous inequality reduces to*

$$Acc(\mathcal{E}) \leq S(\rho).$$

2. *In both cases, for pure states and for mixed states, the equality is attained when the states are mutually orthogonal.*
3. *It can be seen that inequality (3.1) is not tight in general even if we restrict to pure states (see* [**11**, Section 5.4.2]*). In the following, we will see that inequality (3.1) is tight if we allow Alice to use $n$-letters codewords.*

## 4. Classical capacity of a quantum channel

As we explained in Section 4 the classical capacity of a quantum channel is defined as

$$C_c(\mathcal{N}) := \lim_{\epsilon \to 0} \limsup_{k\to\infty} \Big\{ \frac{m}{k} : \exists_{\mathcal{A}}, \exists_{\mathcal{B}} \text{ such that } \|id_{\ell_1^{2^m}} - \mathcal{B} \circ \mathcal{N}^{\otimes_k} \circ \mathcal{A}\| < \epsilon \Big\},$$

where $\mathcal{A} : \ell_1^{2^m} \to \otimes^k S_1^n$ will be a quantum channel representing Alice's encoding from classical information to quantum information and Bob will decode the information he receives from Alice via the $k$ times uses of the channel, $\mathcal{N}^{\otimes_k}$, by means of a quantum channel $\mathcal{B} : \otimes^k S_1^n \to \ell_1^{2^m}$.

As a first approach to study this quantity one could try to follow the proof of the noisy channel coding theorem (see Theorem 3.1 and Subsection 3.1). However, some obstacles appear from the very beginning. A simplification of the problem consists of restricting those encoders used by Alice. Indeed, let us assume that Alice is restricted to the use of those protocols which encode the classical information on elements of the form $\rho_1 \otimes \cdots \otimes \rho_k \in \otimes^k S_1^n$. Then, we talk about the *product state classical capacity of the channel* $\mathcal{N}$ and we denote it by $\chi(\mathcal{N})$.

The same proof of Theorem 3.2 for classical channels applies here to state that

$$C_c(\mathcal{N}) = \sup_k \frac{\chi(\mathcal{N}^{\otimes_k})}{k}. \tag{4.1}$$

The following theorem gives an expression for the product state classical capacity of a quantum channel "analogous" to the expression in the noisy channel coding theorem (Theorem 1.2).

THEOREM 4.1 (Holevo, Schumacher, Westmoreland). *Let $\mathcal{N} : S_1^n \to S_1^n$ be a quantum channel. We have*

$$\chi(\mathcal{N}) = \max_{\{p_j,\rho_j\}} \Big\{ S\Big(\mathcal{N}\big(\sum_j p_j\rho_j\big)\Big) - \sum_j p_j S\big(\mathcal{N}(\rho_j)\big)\Big\},$$

*where the supremum is taken over al ensembles $\{p_j, \rho_j\}$.*

REMARK 4.2. *Note that the previous result states that if we optimize over all possible ensembles $\mathcal{E} = \{p_j, \rho_j\}$ the Holevo information of the ensemble $\mathcal{E}' = \{p_j, \mathcal{N}(\rho_j)\}$ we obtain the product state classical capacity of the quantum channel $\mathcal{N}$. $\chi(\mathcal{N})$ is usually called the Holevo capacity of the cannel.*

HSW Theorem can also be stated in the following way:

THEOREM 4.3. *The classical capacity obtainable using codewords composed of tensor products of signal states $\rho_j$, where the probability of using $\rho_j$ is $p_j$, is given by*

$$\chi(\{p_j, \rho_j\}) = S\big(\sum_j p_j\rho_j\big) - \sum_j p_j S(\rho_j).$$

Indeed, in the case in which Alice is sending information through a quantum channel $\mathcal{N}$, she will consider the ensemble $\mathcal{E} = \{p_j, \rho_j\}$ and we will apply Theorem 4.3 to the ensemble received by Bob $\mathcal{E}' = \{p_j, \mathcal{N}(\rho_j)\}$, to obtain Theorem 4.1. On the other hand, Theorem 4.3 corresponds to the particular case of the identity channel in Theorem 4.1.

REMARK 4.4. *Theorem 4.3 states that the Holevo bound in Theorem 3.1 is attained if we allow Alice to use codewords composed of tensor products of signal states. However, one must be a little bit careful with the meaning of considering the accessible information of an ensemble $\mathcal{E} = \{p_j, \rho_j\}$ when codewords composed of tensor products of signal states are allowed. Indeed, we must understand this capacity $c$ as the possibility of sending $nc$ bits of classical information when we use codewords composed of tensor products of $n$ signal states.*

Note that the statement of Theorem 4.3 is, somehow, more precise than the statement of Theorem 4.1. It states that for a fixed ensemble, $\mathcal{E} = \{p_j, \rho_j\}$, Alice can transmit such an amount of information $(\chi(\{p_j, \rho_j\}))$ to Bob; so the best we can do is to optimize over all possible ensembles. However, we will see that the proof of Theorem 4.1 shows exactly the same for a general channel $\mathcal{N}$.

Lower bound in Theorem 4.1. Let us fix an ensemble $\mathcal{E} = \{p_j, \rho_j\}$ and let us denote $\sigma_j = \mathcal{N}(\rho_j)$ for every $j$. We will use random coding to show that Alice and Bob can communicate $2^{k(R-\delta)}$ bits of classical communication reliably by using $k$ times the channel in parallel and by imposing that our codewords must consist of product states $\rho_{j_1} \otimes \cdots \otimes \rho_{j_k}$. Our goal is to show that $R$ can be taken

$$\chi(\mathcal{E}') = S\big(\sum_j p_j\sigma_j\big) - \sum_j p_j S(\sigma_j),$$

where $\mathcal{E}' = \{p_j, \sigma_j\}$. To each $M \in \{1, \cdots, 2^{kR}\}$ we associate a codeword $\rho_M = \rho_{M_1} \otimes \cdots \otimes \rho_{M_k}$, where each $\rho_{M_i}$ is chosen independently at random from the ensemble $\mathcal{E}$. Let us denote $\sigma_{M_i} = \mathcal{N}(\rho_{M_i})$ for every $i$ and $\sigma_M = \mathcal{N}^{\otimes_k}(\rho_M)$.

We want to construct a POVM $\{E_M\}_{M\cup\{0\}}$ with $E_0 = \mathbb{1} - \sum_{M\neq 0} E_M$, so that the probability of Bob successfully distinguishing the codeword $\rho_M$ is $tr(\rho_M E_M)$ and therefore the probability of error is $P^e_M = 1 - tr(\rho_M E_M)$. Our goal is to prove the existence of a code $\{\rho_M\}_M$ of rate $R > \chi(\mathcal{E}') - \delta$ such that $P^e_M$ is small for every $M$. We will show the existence of such a code verifying that the average error

$$P_{av} = \frac{\sum_M P^e_M}{2^{kR}} = \frac{\sum_M (1 - tr(\rho_M E_M))}{2^{kR}}$$

is small. Then, we can deduce our result by reasoning as we did in the proof of Shannon's noisy channel theorem (see Section 3 in Chapter 6).

We begin by construction the POVM $\{E_M\}_M$. Let $\epsilon > 0$. Then, let us define $\overline{\sigma} = \sum_j p_j\sigma_j$ and let $P$ be the projector onto the $\epsilon$-typical subspace of $\overline{\sigma}^k$. By the theorem of typical subspaces, it follows that for every $\delta > 0$ and sufficiently large $k$ we have

$$tr(\overline{\sigma}^k(\mathbb{1} - P)) \leq \delta. \tag{4.2}$$

For a given $M$, we will also define a notion of an $\epsilon$-typical subspace for $\sigma_M$, based on the idea that typically $\sigma_M$ is a tensor product of $kp_1$ copies of $\sigma_1$, $kp_2$ copies of $\sigma_2$ and so on. Let us define

$$\overline{S} = \sum_j p_j S(\sigma_j).$$

Suppose that $\sigma_j$ has a spectral decomposition $\sum_l \lambda^j_l |e^j_l\rangle\langle e^j_l|$, so that

$$\sigma_M = \sum_L \lambda^M_L |E^M_L\rangle\langle E^M_L|.$$

Here, $L = (l_1, \cdots, l_k)$, and for convenience we define $\lambda^M_L = \lambda^{M_1}_{l_1}\lambda^{M_2}_{l_2}\cdots\lambda^{M_k}_{l_k}$ and $E^M_L = |e^{M_1}_{l_1}\rangle|e^{M_2}_{l_2}\rangle\cdots|e^{M_k}_{l_k}\rangle$. Then, we define $P_M$ as the projector onto the space spanned by all $|E^M_L\rangle$'s such that

$$\Big|\frac{1}{k}\log\frac{1}{\lambda^M_l} - \overline{S}\Big| \leq \epsilon. \tag{4.3}$$

In a similar manner to the proof of the theorem of typical sequences, the law of large numbers implies that for $\delta > 0$ and for sufficiently large $k$ we have

$$\mathbb{E}[tr(\sigma_M P_M)] > 1 - \delta, \tag{4.4}$$

where here the expectation is taken with respect to the distribution over codewords $\rho_M$ (for a fixed $M$) induced by random coding and, thus, for each $M$ we have

$$\mathbb{E}[tr(\sigma_M(\mathbb{1} - P_M))] \leq \delta. \tag{4.5}$$

Also, note that by definition (4.3) the dimension of the subspace onto which $P_M$ projects can be at most $2^{k(\overline{S}+\epsilon)}$ and, thus,

$$\mathbb{E}[tr(P_M)] \leq 2^{k(\overline{S}+\epsilon)}. \tag{4.6}$$

We now use the typicality notion to define Bob's decoding POVM. We define

$$E_M = \Big(\sum_{M'} PP_{M'}P\Big)^{-\frac{1}{2}} PP_MP \Big(\sum_{M'} PP_{M'}P\Big)^{-\frac{1}{2}}.$$

It can be easily seen that $\sum_M E_M \leq \mathbb{1}$ and we can define $E_0 = \mathbb{1} - \sum_M E_M$ to complete the POVM. Note that, up to small corrections, $E_M$ is equal to the projector $P_M$ and Bob's measurement $\{E_M\}$ corresponds essentially to checking if $\rho_M$ falls into the space on which $P_M$ projects.

The technical part of the proof is to show the following upper bound for $P_{av}$, which can be found in [**8**, Box 12.5, page 559].

$$P_{av} \leq \frac{1}{2^{kR}} \sum_M \Big[3tr\big(\sigma_M(\mathbb{1} - P)\big) + \sum_{M \neq M'} tr(P\sigma_M P P_{M'}) + tr(\sigma_M(\mathbb{1} - P_M))\Big]. \tag{4.7}$$

The previous quantity $P_{av}$ is defined with respect to a specific choice of codewords. We are going to calculate the expectation of this quantity over all random codes. By construction, $\mathbb{E}(\sigma) = \overline{\sigma}^k$ and $\sigma_M$ and $P_{M'}$ are independent where $M' \neq M$. Then, we obtain

$$\mathbb{E}[P_{av}] \leq 3tr\big(\overline{\sigma}^k(\mathbb{1} - P)\big) + (2^{kR} - 1)tr(P\overline{\sigma}^k P\mathbb{E}[P_1]) + \mathbb{E}[tr(\sigma_1(\mathbb{1} - P_1))]. \tag{4.8}$$

Using equations (4.2) and (4.5) we can conclude that

$$\mathbb{E}[P_{av}] \leq 4\delta + (2^{kR} - 1)tr(P\overline{\sigma}^k P\mathbb{E}[P_1]).$$

But, $P\overline{\sigma}^k P \leq 2^{-k(S(\overline{\sigma})-\epsilon)}\mathbb{1}$ and, by (4.6) we have

$$\mathbb{E}[tr(P_1)] \leq 2^{k(\overline{S}+\epsilon)},$$

from where we deduce

$$\mathbb{E}[P_{av}] \leq 4\delta + (2^{kR} - 1)2^{-k\left(S(\overline{\sigma})-\overline{S}-2\epsilon\right)}. \tag{4.9}$$

Then, provided $R < S(\overline{\sigma}) - \overline{S}$ it follows that $\lim_k \mathbb{E}[P_{av}] = 0$.

From here we can deduce the existence of a code with rate $R$ by reasoning as in the classical case. □

Upper bound in Theorem 4.1. Let us assume that Alice can reliably send $2^{kR}$ codewords $M$ of the form $\rho_M = \rho_1^M \otimes \cdots \otimes \rho_k^M$ by using $k$ times the channel $\mathcal{N}$ in parallel. Let us denote $\sigma_M = \sigma_1^M \otimes \cdots \otimes \sigma_k^M$ the messages received by Bob and let us assume that Bob's decoder consists of a POVM $\{E_M\}_M$. Without loss of generality we can assume that Bob has one operator $E_M$ for each messages $M$ and possibly an extra element defined as $E_0 = \mathbb{1} - \sum_M E_M$. Therefore, $\{E_M\}_M$ contains at most $2^{kR} + 1$ elements.

As in the proof of the upper bound in Theorem 3.1 we will consider the ensemble formed by our codewords, each with equal probability $2^{-kR}$ and denote by $\tilde{X}^k$ the corresponding random variable. We will also denote by $\overline{X}^k$ the random variable describing the measurement outcome from Bob's decoding. We see that

$$p_{av} := p(\tilde{X}^k \neq \overline{X}^k) = \frac{1}{2^{kR}} \sum_M \big(1 - tr(\sigma_M E_M)\big).$$

According to Fano's inequality we have

$$\begin{aligned} p_{av} kR \geq p_{av} \log(|\tilde{X}^k| - 1) &\geq H(\tilde{X}^k | \overline{X}^k) - H_b(p_{av}) \\ &= H(\tilde{X}^k) - H(\tilde{X}^k : \overline{X}^k) - H_b(p_{av}) \\ &= kR - H(\tilde{X}^k : \overline{X}^k) - H_b(p_{av}). \end{aligned}$$

On the other hand, according to Holevo bound (Theorem 3.1) and the subadditivity of the von Neumann entropy

$$\begin{aligned} H(\tilde{X}^k : \overline{X}^k) &\leq S\big(\sum_M \frac{\sigma^M}{2^{kR}}\big) - \sum_M \frac{S\big(\sigma_1^M \otimes \cdots \otimes \sigma_k^M\big)}{2^{kR}} \\ &\leq \sum_{j=1}^k \Big(S\big(\sum_M \frac{\sigma_j^M}{2^{kR}}\big) - \sum_M \frac{S(\sigma_j^M)}{2^{kR}}\Big). \end{aligned}$$

Since each of these $k$ terms are upper bounded by

$$C := \sup\big\{\chi(\mathcal{E}) : \mathcal{E} = \{p_x, \mathcal{N}(\rho_x)\}\big\},$$

we obtain

$$H(\tilde{X}^k : \overline{X}^k) \leq kC.$$

Hence, we conclude from the previous estimates that $p_{av} kR \geq kR - kC - H_b(p_{av})$, so

$$p_{av} \geq \frac{(R - C)}{R}.$$

Therefore, we must have $R \leq C$. □

## 5. A final comment about the regularization

HSW theorem gives a "simple formula" to compute the product state classical capacity of a quantum channel $\chi(\mathcal{N})$. However, in order to compute the classical capacity of $\mathcal{N}$ we must consider a regularization of this capacity as it was explained in (4.1). A natural question is then if the product state classical capacity of a quantum channel is multiplicative:

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) = \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2),$$

since it would immediately imply that no regularization is required:

$$C_c(\mathcal{N}) = \chi(\mathcal{N}).$$

Recall that we proved in Section 3.1 of Chapter 6 that this is indeed the case when we are dealing with classical channels. The question for quantum channels was a long open problem in the field of quantum information and it was recently solved by Hastings in the negative. Indeed, by means of a probabilistic approach using random unitaries, it was proved in [**7**] the existence of a quantum channel of the form

$$\mathcal{N}(\rho) = \sum_i \lambda_i U_i \rho U_i^* \quad \text{for every} \quad \rho,$$

such that

$$\chi(\mathcal{N} \otimes \overline{\mathcal{N}}) > \chi(\mathcal{N}) + \chi(\overline{\mathcal{N}}).$$

Here, $\{U_i\}_i$ is a family of unitary matrices and $(\lambda_i)_i$ is a probability distribution. Also, the channel $\overline{\mathcal{N}}$ is defined by

$$\overline{\mathcal{N}} = \sum_i \lambda_i U_i^* \rho U_i \quad \text{for every} \quad \rho.$$

By using standard arguments (see for instance [**6**]) one can construct another channel $\Phi$ from $\mathcal{N}$ and $\overline{\mathcal{N}}$ so that

$$\chi(\Phi \otimes \Phi) > 2\chi(\Phi).$$

In particular, Hastings' result tells us that the regularization (4.1) is required for certain channels!

# Bibliography

[1] A. Aspect, P. Grangier, G. Roger, *Experimental Tests of Realistic Local Theories via Bell's Theorem*, Phys. Rev. Lett. 47, 460 (1981).

[2] J. S. Bell, *On the Einstein-Poldolsky-Rosen paradox*, Physics, **1**, 195 (1964).

[3] J. Briët, T. Vidick, *Explicit Lower and Upper Bounds on the Entangled Value of Multiplayer XOR Games*, Comm. Math. Phys. 321 (1), 181-207 (2013).

[4] A. Einstein, B. Podolsky, N. Rosen, *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?*, Phys. Rev. 47, 777 (1935).

[5] A. Grothendieck, *Résumé de la théorie métrique des produits tensoriels topologiques (French)*, Bol. Soc. Mat. Sao Paulo 8, 1-79 (1953).

[6] M. Fukuda, M. M. Wolf, *Simplifying additivity problems using direct sum constructions*, J. Math. Phys. 48, 072101 (2007).

[7] M. B. Hastings, *A Counterexample to Additivity of Minimum Output Entropy*, Nature Physics 5, 255 (2009).

[8] M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, New York, (2000).

[9] V. Paulsen, *Completely bounded maps and operator algebras*, Cambridge Studies in Advanced Mathematics, vol. 78, Cambridge University Press, Cambridge, (2002).

[10] D. Pérez-García, M. M. Wolf, C. Palazuelos, I. Villanueva, M. Junge, *Unbounded violation of tripartite Bell inequalities*, Comm. Math. Phys. 279 (2), 455-486 (2008).

[11] J. Preskill, *Personal notes: http://www.theory.caltech.edu/people/preskill/ph229/notes/chap5.pdf.*

[12] Rowe, M. et al., *Experimental violation of a Bell's inequality with efficient detection*, Nature 409, 791 (2001).

[13] C.E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, 27, 379-423 (1948).

[14] B. S. Tsirelson, *Some results and problems on quantum Bell-type inequalities*, Hadronic J. Supp. 8(4), 329-345 (1993).

[15] Mark M. Wilde, *From Classical to Quantum Shannon Theory*, arXiv:1106.1445.