Essay on the Foundations of Classical and Quantum Information Theories

André Tanus Cesário de Souza

August 2011

Essay on the Foundations of Classical and Quantum Information Theories

André Tanus Cesário de Souza

Orientador:

Prof. Dr. Reinaldo Oliveira Vianna

Versão Final - Dissertação apresentada à UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG, como requisito parcial para a obtenção do grau de MESTRE EM FÍSICA.

Belo Horizonte
Brasil
Agosto de 2011

Dedicate

To my mother Roane Tanus (*in memoriam*).
to my friend Mário Mazzoni,
to Reinaldo O. Vianna,
and also to Karol Życzkowski and Ingemar Bengtsson.

Acknowledgements

Agradeço ao grande professor e físico, meu mestre e amigo *Mário Mazzoni*. Sem a tua amizade e o teu apoio nesse último semestre, seria impossível navegar por esse mar quase infinito de tristezas e dificuldades... muito obrigado pelas lindas poesias e conversas maravilhosas. Agradeço também pelos ensinamentos, tanto no subespaço da técnica, (quase todo o meu conhecimento em física eu aprendi direta ou indiretamente contigo), quanto em todos os outros subespaços da vida.

Ao meu orientador e amigo Reinaldo O. Vianna, pelas infinitas discussões sobre o que há de mais bonito e fundamental na Mecânica Quântica. Seus conselhos sempre foram como uma luz-guia enquanto eu sutilmente divergia.

Agradeço aos professores Aldo Delgado Hidalgo e Carlos Saavedra pelo excelente mês no Chile, na Universidade de Concepción. Um agradecimento especial ao Aldo, pelas discussões e ensinamentos. Agradeço à Telma e Wallon pela imensa e impagável ajuda no Chile. Agradeço os meus amigos Thiago e Wanderson pela maravilhosa viagem Chile afora, por mais de dois mil kilômetros. Sem vocês essa viagem seria impossível!

Agradeço à grande professora Maria Carolina Nemes, (uma das forças da natureza). A imensidão do teu conhecimento ilumina a todos.

Agradeço também aos grandes físicos C. H. Monken e Sebastião de Pádua. Espero aprender muita física com vocês ainda. Agradeço o professor Jafferson Kamphorst, e ainda os professores Marcos Pimenta e João Antônio Plascak pelo exemplo e dedicação, e também os amigos do CETEC-MG: o Reinaldo Trindade, Elenice Cavichioli, Adriano Menezes e o professor J. Roberto T. Branco - pela oportunidade de estudo na Paraíba.

Grande parte do capítulo de formalismo matemático foi inspirado nos dois ótimos livros do professor Marcelo Terra Cunha, portanto o meu obrigado a ele.[1]

Agradeço à minha família: meu Pai, a Raquel e Gustavo (pelos desenhos em Corel) – um beijo para Helena (minha afilhada e paixão da minha vida). Um agradecimento especial à minha avó querida, sempre minha companheira e amiga. Ao meu avô, que sempre foi um exemplo para mim. Ao meu tio Renê -uma das pessoas mais cultas que eu já conheci. Agradeço ao tio Libério que, apesar de não termos contato hoje em dia, grande parte do meu desejo de

[1] I'd like to express my gratitude to Scott Aaronson for the idea followed in the Section 2.6. I'd like also to thank Peter E. Blöchl for the idea followed in Section 4.6.2. This text can be understood as a kind of non-geometrical introduction to the Życzkowski and Bengtsson textbook. A special thanks to the authors mentioned in the bibliography page.

aprender física descende daquelas nossas conversas (juntamente com o meu pai) sobre o universo quando eu era criança. E outro agradecimento especial ao amor da minha vida Ida B. Rodríguez, (e ao Thiaguinho) pela imensa compreensão, amor e carinho. Você é a luz que ilumina o meu universo.

Um agradecimento especial ao grupo InfoQuant: Meu amigo Debarba, (que me emprestou uma fonte de computador para que eu pudesse latecar), meu amigo Fernando, (um grande músico e parceiro), meu amigo e colaborador Thiago, (aprendi muita física contigo) e ao amigo Júnior. Agradeço o Leozão por ter me ensinado muita física. Também agradeço ao grupo Enlight pela enorme colaboração e ajuda dos meus amigos Mateus Araújo e do Marco Túlio. Agradeço aos outros amigos do Enlight: Pierre, o Breno, o Marco Aurélio e o Wanderson, que muito me ensinaram. Agradeço também aos amigos e colegas do Depto. de Física da UFMG[2]: a Amanda e Tauanne pelo carinho, a Lígia pelos sorrisos, a Lídia, Regiane e Jenaína pelas ótimas conversas. Agradeço aos meus amigos: Geraldo Parreiras, a Luciana Cambraia, o Daniel Parreiras, a Ingrid e a Juliana, e o Régis. Agradeço também o meu amigo Tiago Campolina, o Geraldão, a minha amiga Monique, o Daniel Massote, a Fabíola Rocha e os meus grandes amigos Bolivar e Capitão Clyffe. Agradeço também Juan Loredo (meu amigo peruano). Um Agradecimento mais que especial à Nina pelos conselhos e pela amizade: a tua amizade foi e ainda é muito importante para mim, principalmente nos tempos de piração. Aos meus segundo-irmãos Euler P. G. de Mello e Igor Laguna (o meu grande companheiro dos botecos da vida) e Felipe de Lena, que me ajudou na correção do texto. Agradeço à Marluce e Ieda. Um agradecimento especial à Shirley e Maria Clarice. Agradeço ao presidente Luiz Inácio Lula da Silva e à minha querida presidenta Dilma Vana Rousseff pelo melhor governo da História do Brasil. Agradeço também aos Beatles e ao Chico Buarque. E um agradecimento especial ao grande amor da minha vida: Clube Atlético Mineiro.

– The author does not consider this work as completed due to problems after its presentation and lack of review. Suggestions for improvement and error notes can be sent to andretanus@gmail.com
– This text was written by a student for students.
– Este trabalho teve o apoio financeiro direto e indireto da **CAPES**, da **FAPEMIG** e do **CNPq**.

[2]Enumerar pessoas é algo terrível, por isso mantive o critério de ordenamento o mais entrópico possível, com o vínculo da ordem em que os nomes apareciam em minha mente.

Abstract

This work is an essay that assesses some important tools in Classical and Quantum information theory. The concept of qubit (quantum two level systems) is introduced in a simple manner by using a few statistical optics concepts (Stokes vectors and coherence matrices). Unlike most texts of Quantum information in which the qubits of spin are often addressed, in this text the concept of polarization qubit is developed. Shannon and Von Neumann entropy functions are introduced formally and, therefore, some of the particularities of classical and quantum information theory are highlighted and confronted. The most important properties of these entropies are demonstrated with detail. The concept of majorization is introduced for both cases: the quantum and the classical ones. The Jaynes problem is discussed in the light of classical theory of information (The Jaynes principle), in the quantum case, this problem -which is still open- is approached in a simple and intuitive manner.

Keywords

qubit, polarization qubit, Jones vector, Jones matrix, Stokes parameters, coherence matrix, Poincaré–Bloch sphere, convex set, Shannon entropy, conditional entropy, joint entropy, mutual entropy, Von Neumann entropy, relative entropy, entropy properties, entropy inequalities, Kullback-Leibler distance or divergence, majoration, bistochastic matrices, Schrödinger mixture theorem, Jaynes principle, maxent problem, classical information, quantum information.

Resumo

Este trabalho é um ensaio sobre algumas ferramentas importantes para a Informação Clássica e Quântica. O conceito de qubit (sistema de dois níveis quântico) é introduzido de forma simples via óptica estatística (vetores de Stokes e matrizes de coerência). Contrariando a abordagem da maioria dos textos de informação Quântica, nos quais os qubits de spin são frequentemente abordados, nesse texto o conceito de qubit é o de polarização. As noções de entropia de Shannon e de Von Neumann são introduzidas formalmente e, com isso, algumas das particularidades das teorias de informação clássica e quântica são evidenciadas e confrontadas. As propriedades mais importantes dessas entropias são detalhadamente demonstradas. O conceito de majoração é introduzido e desenvolvido em ambos os casos: clássico e quântico. O problema de Jaynes clássico é discutido sob a luz da Teoria de Informação Clássica, (Princípio de Inferência de Jaynes) e, no caso quântico, esse problema -que ainda permanece em aberto- é abordado de forma simples e intuitiva.

Palavras-chave

qubit, qubit de polarização, vetor de Jones, matriz de Jones, parâmetros de Stokes, matriz de coerência, esfera de Poincaré–Bloch, conjuntos convexos, entropia de Shannon, entropia condicional, entropia conjunta, informação mútua, entropia de Von Neumann, entropia relativa, propriedades da entropia, desigualdades para entropia, distância ou divergência de Kullback-Leibler, majoração, matrizes biestocásticas, Teorema da mistura de Schrödinger, princípio de Jaynes, problema maxent, infomação clássica, informação quântica.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This work is organized in seven chapters. Each chapter is systematically organized to contain the many spectrums involved in this project. This chapter, Introduction, briefly puts forward the objective and the scope of study. In Chapter 2, we introduce the polarization qubit (a quantum two level system). We start with the plane-wave solution of the Maxwell's equation for the electric field and observe that all information of the polarization state of a photon can be described as a vector in $\mathcal{C}^2$, called Jones Vector. Any operation in a laboratory can be represented by a matrix called Jones matrix. In a more general situation, the complete description of the polarization state cannot be done by the Jones formalism (by a Jones vector), because both phase and amplitude can vary with time and these fluctuations impose some difficulties. We can associate a set of Jones vectors with a probability vector and apply the concepts of the classical optics with the help of the theory of the probability. But instead, in order to introduce the density matrix formalism for one qubit, we define the coherence matrix, which is an analogue of the density matrix for two level systems.

We understand that we can easily measure time averages in a laboratory and we take advantage of this fact in order to define the coherence matrix by the Stokes parameters. These parameters are defined with an obvious connexion with the Bloch vector representation for one qubit, established with the help of the Pauli matrices. Therefore, the Stokes parameters can be directly related with Bloch vector components. If we allow partial polarization, we can construct a Poincaré sphere and its ball, which forms a complete analogy with the Bloch sphere. Gathering all these coincidences, we finally show an interesting way of constructing a qubit experimentally by using the polarization state of a photon. In summary: if the qubit is fully polarized, then we can use the Jones vector formalism, but if it is partially polarized we will need to define a coherence matrix and this approach follows immediately from the density matrix formalism. Lastly, we discuss some consequences of changing the L_1-norm by the Euclidean L_2-norm. We try to advocate that this *new* theory of probability based on the L_2 norm is a toy-model of the

Quantum Mechanics theory. The essential difference between the qubit and the probabilistic bit is discussed in the light of this theory.

In Chapter 3, we define some mathematical background. As this content might be infinite, we attempt to follow an intuitive ordering and also focus our attention on an axiomatic standpoint. The careful reader can easily notice that this mathematical background does not form a very important section, but its existence is justified by the construction of a *complete* narrative. We almost always try to give more intuitive and easy proofs for all theorems. We begin this chapter with some concepts and definitions about metric spaces and then we try to understand the Quantum Mechanics by using the rules of the C^* non-commutative algebra of its observables. We try to develop a parallel between the quantum and the classical theories. We show that we can understand both theories in an operational way, *i.e.,* first creating an algebra for its observables and later defining the states simply as linear functionals belonging to this algebra.

In chapter 2, we understand in a primary way, some of the similarities of the coherence matrix and the density matrix formalism. We are now then prepared to define mixtures and convex sets more properly. These definitions are important if we want to explore the convexity properties, because the set of the density matrices is a convex set.

The density matrix formalism is formally introduced along with the postulates of Quantum Mechanics. We already discuss the formalism for one particle (one qubit), then for more qubits we need to introduce some properties of the tensor product. The Hilbert-Schmidt's space is also discussed in a brief manner. Finally, we present a very short discussion about the *entanglement* -which is *the* property of Quantum Mechanics. Thus we present an introduction to a geometric interpretation for the entanglement. Despite its importance in Quantum Theory of Information, we do not discuss the entanglement very deeply. In this text, the entanglement is considered only as one of the properties of a quantum multipartite system.

Chapters 4 and 5 are the core of this text. We establish an *invisible* parallel between the Classical and the Quantum Information theories. These chapters abide the following structure: we try to perceive the states as members of a classical and a quantum space of probabilities respectively. The operations allowed in each space are then presented. Some majorization properties are discussed in order to understand the *partial* ordering in these spaces. Finally, we define the entropy function (in the classical case it is the Shannon entropy and in the quantum one the Von Neumann entropy) as a measure of information and a measure of the *degree of mixing*.

In Chapter 4, we introduce the function $\mathcal{I}$ as a measure of *information*, and we define the Shannon entropy as the average missing information, that is, the average information required to specify an outcome when we know its distribution of probabilities. Some important properties of the Shannon entropy are demonstrated with detail. Finally, in Chapter 4, we define the Markovian's processes simply in order to prove some theorems easily. We associate these processes with the bistochastic maps. Later this definition will be useful in order to try to understand the second law of Thermodynamics

for Markovian processes.

In Chapter 5, we present a brief discussion about the stochastic and bistochastic maps (some quantum maps and operations) and about operator ordering. The Von Neumann entropy is defined with rigor and we demonstrate some of its properties. We observe, in both cases, the importance of the relative entropy and we present some of its interesting geometrical properties. In the quantum case, we discuss the mixture problem and the Schrödinger mixture theorem. In both cases, we provide a short discussion on the second law of thermodynamics and we try to exhibit its connexion with the Jaynes principle of maximum entropy. In the classical case, we present other inference schemes. In the quantum one, we present the Jaynes problem in a short, intuitive and easy approach. The reason of this choice of a brief discussion is because this problem is still opened in the quantum case.

CHAPTER 2

What is a Qubit?

2.1 Two Level Systems

The literature of Quantum Mechanics, Quantum Optics and Statistical Mechanics abounds with discussions about *two level states* and *two level systems*. In Statistical Mechanics, for example, a *two level system* can be defined when particles or states can access only two discrete levels with different energies labeled by ε_1 and ε_2. Using the analogy with a classical bit, a quantum bit or simply a *qubit* is also a two level system. These quantum two level systems can be conceived with the spin state of a particle (an electron, for example) [1, 2], or with the polarization state of a photon [1, 3].

In Quantum Information theory, there exists thousands of theorems for systems described by more than one qubit, but how can we create one qubit in a laboratory? In this chapter, we try to answer this question first, by constructing a two level system for pure states of polarization using the *Jones vector formalism* (section 2.2), and later, in the section 2.3, we learn how to proceed operations with these vectors in a laboratory. The analogy between the pure states of polarization, which are completely described by their Jones vector, and the quantum pure vectors is quite obvious and will be addressed. Later, we consider the fact that the Jones vector can vary in time and we develop the *coherence matrix formalism* (section 2.4) to cope with this problem. As this formalism is completely analogous to the formalism of density matrix for one qubit, with the help of these analogies, we can understand the partially polarized states, which are identical entities to the quantum mixed states. In section 2.5, we build the Poincaré–Bloch sphere, that is another consequence of those analogies between the coherence matrix and the density matrix for two level systems. Finally, in the Section 2.6, we try to elucidate what is the essential difference between one qubit and a probabilistic classical bit.

2.2 The Polarization Qubit

A good example for a mathematical analogy of a quantum two level system, *i.e.,* a qubit; is the *polarization* state of a photon, that is a description of the orientation of the oscillations exhibited in transverse waves. We do not observe this property in longitudinal waves because the direction of oscillation and the direction of propagation are the same [4].

Let $\mathbf{E}(\mathbf{r}, t)$ be a *plane-wave* solution for the electric field that travels along the z axis (Eq. 2.1). We can always write the general solution of the wave equation for the electric field (or of the Maxwell's equations in vacuum and without sources) as a linear combination of plane-waves of various frequencies and polarizations, in a linear, homogeneous and time independent media [5] and [6].

$$\mathbf{E}(\mathbf{r}, t) = \begin{pmatrix} E_{0x} e^{i(kz-\omega t+\delta_x)} \\ E_{0y} e^{i(kz-\omega t+\delta_y)} \\ 0 \end{pmatrix}, \tag{2.1}$$

Where δ_x and δ_y are phases, $E_{0x} = |\mathbf{E}|\cos\theta$, $E_{0y} = |\mathbf{E}|\sin\theta$, the angle of polarization θ is given by $\theta = \arctan(\frac{E_{0y}}{E_{0x}})$ and $|\mathbf{E}|^2 = E_{0x}^2 + E_{0y}^2$. With the electric field in hands, we can obtain the magnetic field easily by doing the following cross product $\mathbf{B}(\mathbf{r}, t) = \frac{\hat{z}\times\mathbf{E}(\mathbf{r},t)}{c}$.

Since only measure light intensities, which are proportional to the squared electric field, ($I \propto |\mathbf{E}|^2$), the global phase (we define the phase $e^{i\delta_x}$ in Eq. 2.2 as the global phase), does not have any physical meaning and only the relative phase $e^{i(\delta_y-\delta_x)}$ has importance, then we can write [4]:

$$\mathbf{E}(\mathbf{r}, t) = |\mathbf{E}| \begin{pmatrix} \cos\theta \\ \sin\theta e^{i(\delta_y-\delta_x)} \\ 0 \end{pmatrix} e^{i\delta_x} e^{i(kz-\omega t)}. \tag{2.2}$$

We can define completely the polarization state of a photon using this monochromatic plane-wave solution (without sources) shown in Eqs. 2.1 and 2.2. Because the electric field wave is traveling in z direction (by definition), all the information of a polarization state of a photon can be written as a vector in $\mathcal{C}^2$, in the xy plane. This vector (Eq. 2.3) is known as *Jones vector*[1] [5]. Note that the oscillatory term $e^{i(kz-\omega t)}$ is from now on omitted.

$$|E\rangle = \begin{pmatrix} E_{0x} e^{i\delta_x} \\ E_{0y} e^{i\delta_y} \end{pmatrix}. \tag{2.3}$$

This general Jones vector shown in Eq. 2.3, *i.e.* $|E\rangle$, is *elliptically* polarized. We say that a wave is *linearly* polarized at an angle θ *i.e.*, $|\theta\rangle$, if $E_{0x} = |\mathbf{E}|\cos\theta$, $E_{0y} = |\mathbf{E}|\sin\theta$, and $\delta_x = \delta_y = \delta$. If $\theta = 0$ in Eq. 2.3, we have *horizontal* polarization $|H\rangle$, and if $\theta = \frac{\pi}{2}$, then we will have *vertical* polarization $|V\rangle$,

$$|\theta\rangle = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \tag{2.4}$$

[1] Be the reader aware that we are going to use the Dirac notation everywhere in this dissertation, not only for quantum states, then $\mathbf{E}(\mathbf{r}) \equiv \vec{E}(\vec{r}) \equiv |E(\mathbf{r})\rangle \equiv |E\rangle$.

(Here we made $|\mathbf{E}|= 1$). We will have circular polarization ($|c\rangle$) if $\theta = \frac{\pi}{4}$ and if $\delta_y = \frac{\pi}{2} \pm \delta_x$ in Eq. 2.3:

$$|c\rangle = \frac{|\mathbf{E}|}{\sqrt{2}} \begin{pmatrix} 1 \\ \pm i \end{pmatrix}. \tag{2.5}$$

The electric field vector in Eq. 2.5 describes a circle of radius equal to $|\mathbf{E}|$ in a perpendicular plane to the z axis. This is a plane wave *circularly* polarized, this vector indicates that $\mathbf{E}$, from plane to plane, has a constant modulus while its direction constantly rotates. Circular polarization may be referred to as right-handed $|c\rangle = |R\rangle$, or left-handed $|c\rangle = |L\rangle$, depending on the direction in which the vector $\mathbf{E}$ rotates. Unfortunately there are two historical conventions still used in the literature.

2.3 Jones Matrices

As discussed in the last section, the complete information of the amplitude and relative phase can be represented in a two dimensional complex vector $|E\rangle$ called *Jones vector*. We represent an operation in a laboratory, such as projections, wave plates, *etc.* by a 2×2 complex matrix L called *Jones matrix*[2], see for example [5, 7]. This matrix is defined in Eq. 2.7.

$$\left|\tilde{E}\right\rangle = L|E\rangle, \tag{2.6}$$

$$L = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{2.7}$$

Where a, b, c and d are complex numbers [7]. We can represent n operations in a laboratory by a product of its Jones matrices, *i.e.*, $\left|\tilde{E}\right\rangle = L_n \cdots L_2 L_1 |E\rangle$. A good example of a Jones matrix is a *projector*. Let us define the projectors as a rank one and trace one Hermitian matrix $|\psi\rangle\langle\psi| \equiv P_\psi$ that satisfies the following property

$$P_\psi^n = P_\psi, \quad \forall n \in \mathcal{N}. \tag{2.8}$$

Using the basis $\{|H\rangle = (1, 0)^\dagger,\ |V\rangle = (0, 1)^\dagger\}$, the projectors P_H and P_V are:

$$P_H = |H\rangle\langle H| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.9}$$

$$P_V = |V\rangle\langle V| = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.10}$$

These two projectors form a resolution of the identity operator[3], granting first that those projectors be normalized.

[2] These matrices can be projections, then they do not represent in general *trace preserving* operations.

[3] In a two dimensional space, the summation over any two orthogonal projectors forms a resolution of the identity operator, *i.e.*, $|{+45°}\rangle\langle{+45°}| + |{+135°}\rangle\langle{+135°}| = |R\rangle\langle R| + |L\rangle\langle L| = |H\rangle\langle H| + |V\rangle\langle V| = \mathbb{I}$. The reader can confirm this fact by performing these summations with the projectors given in the table of subsection 2.5.1

2.4 Coherence Matrix Formalism

In a more general situation, the complete description of the polarization state cannot be done by the Jones formalism because both phase and amplitude vary with time and these fluctuations impose some difficulties. Then the vector shown in Eq. 2.3 depends on time in a explicit manner, as shown in Eq. 2.11. A convenient way to supplant this difficulty is by using the *coherence matrix formalism* [7]. We could associate a set of Jones vectors with a probability vector depending on time and apply the concepts of the classical optics with the help of the probability theory. But in order to introduce later the density matrix formalism for one qubit, we will define the coherence matrix, which is an analogue of the density matrix for two level systems. In the Chapter 3, Section 3.4, the real difficulty of the "braket-vector formalism" is clarified and we will introduce the "density matrix formalism" more rigorously. For this Chapter purposes, it is sufficient to define the coherence matrix and the density matrix and show the obvious analogies between these two entities.

$$|E(t)\rangle = \begin{pmatrix} E_{0x}(t)e^{i\delta_x(t)} \\ E_{0y}(t)e^{i\delta_y(t)} \end{pmatrix}. \tag{2.11}$$

Let us define a positive semi-definite matrix J called *coherence matrix*:

$$J = \begin{pmatrix} \langle E_x(t)E_x(t)^* \rangle & \langle E_x(t)E_y(t)^* \rangle \\ \langle E_y(t)E_x(t)^* \rangle & \langle E_y(t)E_y(t)^* \rangle \end{pmatrix}, \tag{2.12}$$

Where $\langle\ \rangle$ means time average. As the laboratory instruments can measure these averages, then it is useful to define such matrix. The total intensity of the field is given by the trace of the coherence matrix, *i.e.;* $I_0 = Tr(J) = |E_x|^2 + |E_y|^2$. As discussed above, the linear operations performed in a laboratory can be mapped in a Jones matrix L defined in Eq. 2.6 and 2.7. Then using Eq. 2.6, 2.11 and 2.12 and the fact that $(AB)^\dagger = B^\dagger A^\dagger$, we can construct a rule of transformation from the coherence matrix J to another one called $\tilde{J}$ with the help of the Jones matrices (L), [7]:

$$\tilde{J} = \left|\tilde{E}\right\rangle\left\langle\tilde{E}^\dagger\right| = L\left|E\right\rangle\left\langle E^\dagger\right|L^\dagger, \tag{2.13}$$

$$\tilde{J} = LJL^\dagger. \tag{2.14}$$

2.5 The Stokes Parameters and Poincaré-Bloch Spheres

2.5.1 The Stokes Parameters

The generic state of a qubit can be specified by a real vector, called *Stokes vector*. The components of this four dimensional vector, the *Stokes parameters*, have the advantage of directly corresponding to empirical quantities, such as photon-counting rates [8]. These parameters will be useful later in order to define the *Poincaré–Bloch spheres*. First let us define the *Pauli matrices* as 2×2 Hermitian matrices σ_i with $i = 1, 2, 3$ (Eqs. 2.15, 2.16 and 2.17). The Pauli

matrices[4] have trace zero and eigenvalues equal to ± 1. The identity matrix is defined in Eq. 2.18.

$$\sigma_1 = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{2.15}$$

$$\sigma_2 = \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{2.16}$$

$$\sigma_3 = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.17}$$

$$\sigma_0 = \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.18}$$

The Pauli matrices and the identity matrix form a complete basis for the space of 2×2 Hermitian matrices. The coherence matrix can be written in such basis if we redefine the Pauli matrices and the identity matrix as $X_i = \frac{\sigma_i}{2}$, $\forall i = 0, 1, 2, 3$ [9]. By using this definition[5], we can expand the matrix J as:

$$J = \sum_{i=0}^{3} S_i X_i. \tag{2.19}$$

The S_i parameters in the "Pauli" X_i basis are called the *Stokes parameters* [8, 9]. The matrix J shown in Eq. 2.20 can be written explicitly as:

$$J = \frac{1}{2} \begin{pmatrix} S_0 + S_3 & S_1 - iS_2 \\ S_1 + iS_2 & S_0 - S_3 \end{pmatrix}. \tag{2.20}$$

This matrix J must be positive semidefinite and Hermitian, and it is indeed. The Pauli matrices form an orthonormal basis[6] for the 2×2 Hermitian matrices. If we want to calculate the *ith*-Stokes parameter we just need to apply the rule $S_i = Tr(JX_i)$ or compare the Eq.s 2.12 and 2.19 [7–9]. By using the latter rule we can calculate the Stokes parameters directly [6]:

$$S_0 = \left\langle |E_x|^2 + |E_y|^2 \right\rangle, \tag{2.21}$$

$$S_1 = \left\langle E_x E_y^* + E_x^* E_y \right\rangle, \tag{2.22}$$

$$S_2 = i\left\langle E_x E_y^* - E_x^* E_y \right\rangle, \tag{2.23}$$

$$S_3 = \left\langle |E_x|^2 - |E_y|^2 \right\rangle. \tag{2.24}$$

[4] The non-trivial products of the Pauli matrices are given by $\sigma_i \sigma_j = \delta_{ij}\sigma_0 + i\varepsilon_{ijk}\sigma_k$.

[5] In H. Pires dissertation, [9], we found another definition for these matrices: $X_i = \frac{\sigma_i}{\sqrt{2}}$.

[6] With respect to the inner product $Tr(\sigma_i \sigma_j)$.

The physical interpretation of this parameters is shown below [6]:

$$S_0 = \langle I_0 \rangle, \tag{2.25}$$
$$S_1 = \langle T_{+45^\circ} - T_{+135^\circ} \rangle, \tag{2.26}$$
$$S_2 = \langle T_R - T_L \rangle, \tag{2.27}$$
$$S_3 = \langle T_H - T_V \rangle. \tag{2.28}$$

Eq. 2.25 shows that the number S_0 is the total average intensity I_0 [9]. Eq. 2.26 relates the difference of the intensities at 45° and 135° polarizations with S_1, Eq. 2.27 determines the difference of the intensities at right and left circular polarizations (R, L) with S_2, and Eq. 2.28 assigns the difference of the intensities at horizontal and vertical polarizations (H, V) with S_3 [9].

2.5.2 The Poincaré–Bloch Spheres

It can be shown [7] that in the Stokes representation for the electric field, the polarization degree P, which is a number $0 \leq P \leq 1$, is given by Eq. 2.29. This number says if the wave is partially or completely polarized.

$$P = \frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0}. \tag{2.29}$$

If $P = 1$, we have a fully polarized wave or photon. In this case, Eq. 2.29 is an equation of a sphere in the Stokes space (S_1, S_2, S_3) with radius $r = PS_0$ [3, 8]. The sphere which equation is shown in Eq. 2.30 is called the *Poincaré sphere* [3, 7].

$$S_1^2 + S_2^2 + S_3^2 = (PS_0)^2. \tag{2.30}$$

It is easy to see in Eq. 2.29 that for partially polarized states, *i.e.*, $0 \leq P < 1$, the radius of the Poincaré ball will be $r = PS_0$. If we consider unpolarized radiation, ($P = 0$), the radius of the ball will be $r = 0$, *i.e.*, $S_1^2 + S_2^2 + S_3^2 = 0$. Each point in this sphere defines a state of polarization which can be completely determined if we measure all parameters needed, or in other words, if we perform a quantum state *tomography*. For a simple introduction of this subject, see for example, [10] and for more involved projects see [11].

If we normalize the matrix J by its trace, ($Tr(J) = S_0$), we can create a matrix ρ that is an analogy of the *density matrix* for one qubit.

$$\rho = \frac{J}{Tr(J)}, \tag{2.31}$$
$$\rho = \frac{1}{2}\begin{pmatrix} 1 + \frac{S_3}{S_0} & \frac{S_1 - iS_2}{S_0} \\ \frac{S_1 + iS_2}{S_0} & 1 - \frac{S_3}{S_0} \end{pmatrix}. \tag{2.32}$$

In the table below, (Table 2.1), we write the Jones vector $|E\rangle$, its coherence matrix J and the Stokes vector $S = (S_0, S_1, S_2, S_3)$ for some pure states of polarization: (H, V, +45°, +135°, R, L). This table can be found in [7] and it is not normalized just for aesthetic reasons.

	$\|E\rangle$	J	S
H	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\\0\end{pmatrix}$
V	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}$	$\begin{pmatrix}1\\-1\\0\\0\end{pmatrix}$
$+45°$	$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\\0\end{pmatrix}$
$+135°$	$\begin{pmatrix}1\\-1\end{pmatrix}$	$\begin{pmatrix}1 & -1\\-1 & 1\end{pmatrix}$	$\begin{pmatrix}1\\0\\-1\\0\end{pmatrix}$
R	$\begin{pmatrix}1\\-i\end{pmatrix}$	$\begin{pmatrix}1 & i\\-i & 1\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\\1\end{pmatrix}$
L	$\begin{pmatrix}1\\i\end{pmatrix}$	$\begin{pmatrix}1 & -i\\i & 1\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\\-1\end{pmatrix}$

Table 2.1: The Jones vector, its coherence matrix and the Stokes vector.

Let us define new parameters $x = \frac{S_1}{S_0}$, $y = \frac{S_2}{S_0}$ and $z = \frac{S_3}{S_0}$ and a vector $\mathbf{r}$ defined as $\mathbf{r} = (r_1, r_2, r_3)^\dagger = (x, y, z)^\dagger$. Then $\rho = \frac{1}{2}(\mathbb{I} + \sum_i r_i \sigma_i)$. Thus, another way to write this expansion is:

$$\rho = \frac{1}{2}\begin{pmatrix}1+z & x-iy\\ x+iy & 1-z\end{pmatrix}, \tag{2.33}$$

$$\rho = \frac{\mathbb{I} + \mathbf{r}\cdot\sigma}{2}. \tag{2.34}$$

Any arbitrary trace one 2×2 Hermitian matrix can be parameterized as shown in Eq. 2.34 [12]. The matrix ρ in this equation is exactly the definition of the *density matrix* for one qubit [10]. The vector $\mathbf{r} = (x, y, z)$ is called *Bloch vector* and its components are the coordinates of the Pauli matrices $(\sigma_x, \sigma_y, \sigma_z)$ space [10, 13]. This matrix is positive-semidefinite, (*i.e.,* its eigenvalues are non-negative), *iff* the parameters x, y, z satisfies the inequality given in Eq. 2.35

$$x^2 + y^2 + z^2 \leq 1 \tag{2.35}$$

In the Poincaré sphere the coordinates are the Stokes parameters, which are related to the Bloch sphere by the Pauli matrices coordinates. As the Poincaré sphere is a sphere in the space of the coherence matrices, the Bloch sphere

is also a sphere in the space of the density matrices, see Fig. 2.1. Using the analogy with the Poincaré sphere shown in Eq. 2.30, we can define another sphere and its ball with the equation given in Eq. 2.35. They are called *Bloch sphere and Bloch ball* respectively and they correspond to the space of the density matrices for one qubit, pure and mixed respectively[7] [4].

The Fig. 2.1, is a representation of the *Poincaré–Bloch sphere*.

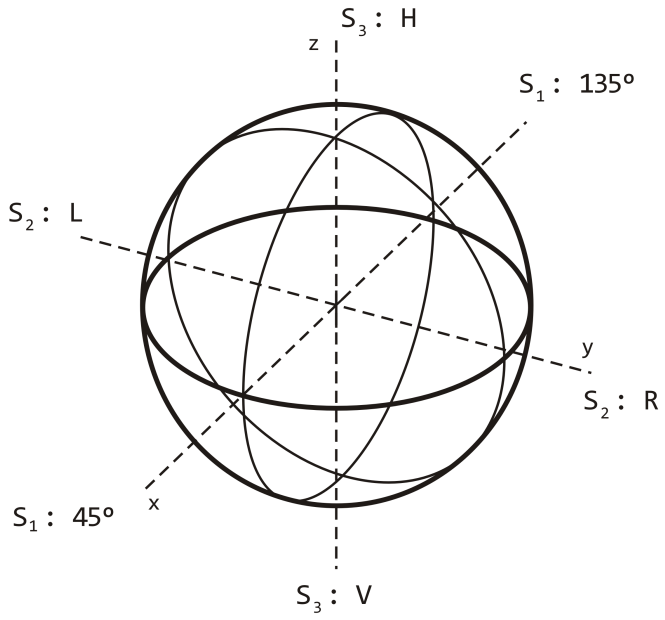

Figure 2.1: The Poincaré–Bloch Sphere.

If $|\mathbf{r}| = 1$, then we have a pure state and it is a point in the Bloch sphere [14]. This quantum pure state described by a trace one Hermitian matrix as in Eqs. 2.33 or 2.34 is a complete analogy of a completely polarized state, described by the matrix J shown in Eqs. 2.19 or 2.20 *i.e.,* with $P = 1$.

If $0 \leq |\mathbf{r}| < 1$ we have a mixed state which is a point inside the Bloch sphere, or a point in the Bloch ball. This is also an analogy with a partial polarized state, which is a point in the Poincaré ball with $r = PS_0$ and $0 \leq P < 1$. For a completely unpolarized state, the matrix in Eq. 2.20 is a multiple of the identity matrix, $J = \frac{S_0}{2}\mathbb{I}$, but $S_0 = I_0$, then $J = I_0\frac{\mathbb{I}}{2}$. In the Bloch ball, this state occurs when $|\mathbf{r}| = 0$, and the matrix shown in Eqs. 2.33 and 2.34 is the trace one identity matrix ($\frac{\mathbb{I}}{2}$).

Theorem 2.1. *A quantum state ρ is a pure state iff $Tr(\rho^2) = 1$. For one qubit this condition implies that $|\mathbf{r}| = 1$.*

[7] In higher dimension, *i.e.,* (higher than the one qubit space studied here in this chapter $Dim = 2 \times 2$), the geometry of the quantum states is much more complicated.

Proof:[8] The proof can be done in two ways: we can square the matrix ρ and imply the trace condition $Tr(\rho^2) = 1$ or we can expand the matrix ρ in Pauli matrices and use their orthogonality properties. The first way gives us the following equality $Tr(\rho^2) = \frac{1}{2}(1 + |\mathbf{r}|^2)$. Since the positivity of ρ implies that $0 \leq |\mathbf{r}| \leq 1$ (Eq. 2.35, then $Tr(\rho^2) = 1$ *iff* $|\mathbf{r}| = 1$. This also implies that a pure qubit lies somewhere on the Bloch sphere, *i.e.*, $x^2 + y^2 + z^2 = 1$. □

Theorem 2.2. *A polarization state described by a coherence matrix J is a pure state (or it is fully polarized) iff $Tr(J^2) = S_0^2$. This condition implies that $S_1^2 + S_2^2 + S_3^2 = S_0^2$, i.e., it lies on the Poincaré sphere.*

Proof: The proof is trivial since we have already showed that $\rho^2 = \frac{1}{S_0^2} J^2$. □

Remark: Note that we are not interested in the spherical representation of the states in both spheres. We do not discuss about the angles which define a state in each sphere: $J(PS_0, \theta_S, \phi_S)$ and $\rho(|\mathbf{r}|, \theta_B, \phi_B)$. The angle subtended by a pair of directions in Hilbert space is *half* the corresponding angle in the Poincaré sphere, which corresponds to the fact that the group $SU(2)$ acting on the vectors in the complex representation is the universal double covering group of rotations $SO(3)$ of vectors in real representation, see [8].

2.5.3 Experimental Determination of One Qubit

Let us consider the generic case of a state with an unknown preparation. Of course we allow partial polarization. So the problem is: how can we determine the density matrix ρ or equivalently determine the correspondent coherence matrix J? We need to measure some observables. Measuring the vertical-horizontal polarization is equivalent to measure an observable σ_z (Eq. 2.17). Its eigenvectors are the pure states $\{|H\rangle, |V\rangle\}$, corresponding to eigenvalues equal to ± 1. Likewise a test for the $\pm 45°$, corresponding to the pure states $\{|\pm 45°\rangle\}$, is equivalent to measure the observable σ_x (Eq. 2.15), and a test for left and right circular polarization that corresponds to the pure states $\{|R\rangle, |L\rangle\}$, is equivalent to measure the observable σ_y (Eq. 2.15). These three measurements, and the total intensity (it corresponds to the normalization factor, *i.e.*, the identity matrix), repeated many times on three disjoints sets of a light beam yield the following set of averages: $r_i = \langle \sigma_i \rangle = Tr(\rho\sigma_i) = Tr(J\sigma_i)$ [15]. The observed values for the Bloch-Poincaré vector $\mathbf{r}$ allow us to write the density matrix $\rho = \frac{1}{2}(\mathbb{I} + \mathbf{r} \cdot \sigma)$. Another approach for an experimental method of determination of one qubit can be seen in [10], see chapter 8.

[8] The word *proof* have distinct meanings in this text. Some proofs are out of the scope of this work because they are much difficult or too large or they use a very strict mathematics, (remember that this is a physicist's work, written for physicists). Sometimes we try to give simply an idea or sketch of the proof and sometimes this word will means its mathematical strict meaning. The distinctions between these meanings is obvious by the rigor adopted in each case.

2.6 What is the Essential Difference Between one Qubit and a Probabilistic Classical Bit?

Of course this discussion could be infinite. For example, if we have more than one qubit, in the quantum case, we may have entanglement which is a property that does not even exists in the classical counterpart. We do not walk toward this line in this section. The idea here is trying to elucidate the essential difference between one qubit and a probabilistic classical bit. But why probabilistic bit? The answer is quite simple because the ordinary deterministic classical bit is less interesting and it does not have any of the following properties discussed below[9].

2.6.1 "Negative Probabilities" - Interference

> Quantum mechanics permits the cancellation of possibilities.
>
> Nick Herbert, Quantum Reality.

Let us suppose that an event having n different outcomes such that each outcome can be associated with a probability p_n of occurrence. We can construct a vector of probabilities with this set of n probabilities. By the classical theory of probability we know that this vector $\vec{p} = (p_1, p_2, \cdots, p_n)^\dagger$ must be positive, and normalized *i.e*, $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$. We can express all these facts above by saying that the L_1-norm of the probability vector is equal to one, that is $\|\vec{p}\|_1 = \sum_i p_i = 1$. As there exists many other norms defined in the metric spaces, (see Chapter 3 for a naive and brief discussion about norms in metric spaces), we could use any other norm to measure the *length* of this probability vector. If we try to use the Pythagorean norm, *i.e.*, the L_2-norm, we would need to take the square root of the sum of the squares of the entries, or $\|\vec{p}\|_2 = \sqrt{\sum_i p_i^2} = 1$ [16].

What are the consequences if we try to construct another theory of probability based on the L_2-norm instead of the L_1-norm? As S. Aaronson made in [16], we try to convince you that quantum mechanics is what necessarily results. We know that the building block of the classical information theory is the *bit*, which can assume the value 0 or 1. In the classical probability theory, we define a probabilistic classical bit when we associate a probability p of occurrence of the value zero and a probability $(1 - p)$ of being one. If we change the L_1-norm by the L_2-norm, and keep the focus on the *real numbers*[10], we will need numbers such as their square add to 1. The set of *all* binary vectors $\vec{p} = (\alpha, \beta)^\dagger$ such as $|\alpha|^2 + |\beta|^2 = \alpha^2 + \beta^2 = 1$ describes a circle of radius equal to 1 [16].

Let us suppose then that we have a binary vector $\vec{p} = (\alpha, \beta)^\dagger$ which is described by this new theory of probability based on the L_2-norm. We cannot

[9] This section was totally inspired in this wonderful lecture: [16].

[10] Are the complex numbers really needed in this "L_2-norm theory of probability"? Yes. But Why? The reason is because the amplitudes in quantum mechanics are given by complex numbers.

assign directly the probability of performing an experiment and obtain a result 0, for example. We need to define something like a *probability amplitude*, given by α and β, in the binary case, such that the probability of obtaining an outcome is just the square of its amplitude, *i.e.*, $p(0) = \alpha^2$ and $p(1) = \beta^2$, with $p(0) + p(1) = 1$. If we try to use the old L_1-norm to measure this vector you will find that $\|\vec{p}\|_1 = \alpha + \beta \neq 1$. But if we use the L_2-norm we will find $\|\vec{p}\|_2 = \sqrt{\alpha^2 + \beta^2} = \sqrt{1} = 1$.

Then why not forget the amplitudes α and β and describe the bit just by its probabilities? The difference is in how the vector is transformed when we perform an operation on it. Indeed, a bistochastic matrix is the most general matrix that always maps a probability vector to another probability vector [16], *i.e.*, a bistochastic map preserves the L_1-norm. But in the L_2-norm, the most general matrix that always maps a normalized vector in the L_2-norm to another normalized vector in the L_2-norm is the unitary matrix. A unitary matrix[11] U is a Hermitian matrix such that $U^\dagger = U^{-1}$, or $U^\dagger U = \mathbb{I}$.

We know that the quantum *bit*, *i.e.*, the *qubit* is the counterpart of the classical bit. Let be $\{|0\rangle, |1\rangle\}$ a basis for the two-dimensional Hilbert space. Then any qubit can be written in such basis. The most general qubit can be expressed as: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, with $\alpha, \beta \in \mathcal{C}$, and $|\alpha|^2 + |\beta|^2 = 1$. As discussed in this chapter, the two labels "0" and "1" written in this qubit could represent any orthogonal basis of the Hilbert space. It could represent, for example vertical and horizontal polarization, that is $|\psi\rangle = \alpha|H\rangle + \beta|V\rangle$.

Let us suppose that we have a qubit ket-vector given by $|\psi\rangle = (\alpha, \beta)^\dagger$, measured in the L_2 norm, *i.e.*, $\alpha^2 + \beta^2 = 1$, and let us also suppose, for the sake of simplicity, that all the amplitude of probabilities are given by real numbers. Then we will obtain another vector $|\psi'\rangle$ if we act a unitary matrix U in the previous vector, as described in Eqs. 2.36 and 2.37.

$$|\psi'\rangle = U|\psi\rangle, \tag{2.36}$$

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ -e^{i\theta}u_{12}^* & e^{i\theta}u_{11}^* \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{2.37}$$

Were u_{11} and u_{12} are complex numbers such that $|u_{11}|^2 + |u_{12}|^2 = 1$. The unitary matrix U given in Eq. 2.37 is the general expression of a unitary matrix that acts on $\mathcal{C}^2$, with $\det(U) = e^{i\theta}$. It depends on four Real parameters: the phase of u_{11} and u_{12}, the phase θ and the relative magnitude between u_{11} and u_{12}.

An Example

Now that we know that we can transform any qubit by applying a 2×2 unitary matrix, let us consider, as an example, a quantum state initially in the state $|0\rangle$, and the following unitary matrix U, written in the basis (Eq. 2.38), $\{|0\rangle, |1\rangle\} = \{|0\rangle \equiv (1, 0)^\dagger$ and $|1\rangle \equiv (0, 1)^\dagger\}$. The task here is evolve

[11] See Chapter 3.

this quantum pure state by acting twice this unitary U.

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. \tag{2.38}$$

As shown in the Fig. 2.2, we apply this unitary transformation U twice in our qubit, which is initially in the state $|0\rangle$, and try to interpret the result:

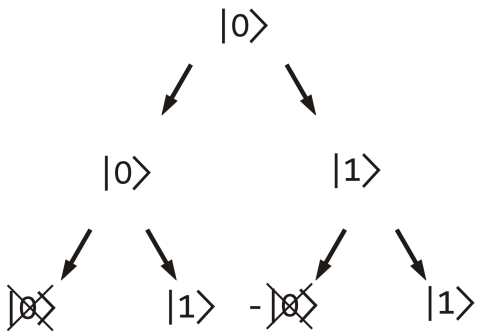

Figure 2.2: An example of one qubit unitary evolution.

As discussed before, let us suppose that our qubit is initially in the state $|\psi\rangle = |0\rangle = (1, 0)^{\dagger}$ (see Fig. 2.2). If we apply the unitary matrix U given by Eq. 2.38 to $|0\rangle$, we will find the state $\frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]$, in other words: $U|0\rangle = \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]$, with probability of obtaining the system in the state $|0\rangle$ given by $p(0) = \left|\frac{1}{\sqrt{2}}\right|^2 = \frac{1}{2}$ and with probability of seeing the system in the state $|1\rangle$ given by $p(1) = \left|\frac{1}{\sqrt{2}}\right|^2 = \frac{1}{2}$, remember that in this new norm the amplitude of probabilities does not sum to 1. For this reason, we need to define a vector with the amplitude of probabilities and also define the probability as the square of the absolute value of this amplitude. Thus, this vector can be now defined as $\vec{p} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, which gives us a state vector written as $\frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]$. But if we apply the matrix U once again we will obtain $|1\rangle$ with 100% of certainty!

As we could think the first operation U as a coin tossing, after the first operation we had two different *paths* to follow: the *path* "0" and "1", each one being equally likely. But if we apply a *randomizing* operation twice to a *random* state, we will produce a deterministic outcome $|1\rangle$! This is the so-called phenomenon of *interference*. The non-observed paths that lead to the outcome zero "0" interfere destructively and cancel each other out. However,

the two paths leading to the outcome $|1\rangle$, both possess positive amplitude, and therefore, they interfere constructively [16].

This fact was also observed in Quantum Random Walk theory. The Quantum Random walk was firstly discussed by Y. Aharonov, L. Davidovich and N. Zagury in 1993 in [17]. In 2003 J. Kempe [18] used a *Hadamard* coin H, (a unitary matrix), as defined in Section 5.3.3, and performed a quantum random walk simulation obtaining the same result of path interference, which can be seen in a clear asymmetry between left and right. The conclusion of Kempe was: "A first thing to notice is that the quantum random walk induces an asymmetric probability distribution on the positions, it is *drifting* to the left. This asymmetry arises from the fact that the H coin (a unitary matrix) treats the two directions $|\leftarrow\rangle$ and $|\rightarrow\rangle$ differently; it multiplies the phase by -1 only in the case of $|\rightarrow\rangle$. Intuitively this induces more cancellations for paths going right-wards (destructive interference), whereas particles moving to the left interfere constructively" [18]. Of course you never see this kind of interference in the classical world. The reason is because in our classical world the probabilities cannot be negative. This phenomenon of *interference* caused by the cancellation between positive and negative amplitudes of probability can be seen as the source of some of the *quantum weirdness* [16].

CHAPTER 3

Mathematical Background–The Space of Density Matrices

In this chapter, we try to define some mathematical background. As this content might be infinite, we attempt to follow an intuitive ordering and also focus our attention on an axiomatic standpoint.

3.1 Some Important Facts About Metric Spaces

3.1.1 The Distance Function

In order to distinguish elements of a set, we require the concept of distance. In metric spaces there exists our intuitive concept of *distance*. A metric space is a pair $(\mathcal{V}, d)$, where $\mathcal{V}$ is a set and d is a distance function[1], such that $d(\star, \star) : \mathcal{V} \times \mathcal{V} \longmapsto \mathcal{R}^+$ [19]. This function d is called *metric* or *distance function*. Let $x, y, z \in \mathcal{V}$, in order to d be a distance function, or a metric, the following properties must hold:

1. $d(x, y) \geq 0, \ \forall x, y \in \mathcal{V}$,
2. $d(x, y) = 0$, *iff* $x = y$,
3. $d(x, y) = d(y, x)$, and
4. $d(x, y) \leq d(x, z) + d(y, z), \ \forall x, y, z \in \mathcal{V}$.

Definition 3.1. *Let x and y be two points of $\mathcal{R}^n$. Then the L_p-distance is defined as:* $d_p(x, y) = (\sum_i |x_i - y_i|^p)^{\frac{1}{p}}$. *If* $p \to \infty$, $d_\infty(x, y) = max_{\{1 \leq i \leq n\}}(|x_i - y_i|)$.

[1] The $\star$ represents any mathematical entity or mathematical element. Then $f(\star)$ represents an empty function such as $f(\star) = f(x)$, if $x = \star$.

Definition 3.2. *The distance between two points in the Euclidean space $\mathcal{R}^n$ is given by the L_2-distance: $d_2(x, y) = (\sum_i |x_i - y_i|^2)^{\frac{1}{2}}$.*

3.1.2 Norms

A norm is a function $\|\star\| : \mathcal{V} \longmapsto \mathcal{R}^+$ that associates each vector $x \in \mathcal{V}$ to the real number $\|x\|$, called the norm of x [19]. If a metric function d defined on a vector space $\mathcal{V}$ satisfies the translational invariance property $d(x - u, y - u) = d(x, y)$ for any vector x, y, and u, and the scale invariance property[2] $d(\beta x, \beta y) = |\beta| d(x, y)$, then we can define a *norm* on the set $\mathcal{V}$ as $\|x\| \equiv d(x, 0)$. Every normed space becomes a metric space if we define the metric $d(x, y) \equiv \|x - y\|$. This metric is called *induced* by the norm [19]. If $\|\star\|$ is a norm, then for all $x, y \in \mathcal{V}$ and $\lambda \in \mathcal{C}$, then all these following conditions must hold:

1. $\|x\| = 0$, *iff* $x = 0$,
2. $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$,
3. $\|x + y\| \leq \|x\| + \|y\|$.

Definition 3.3 (Sup-norm). $\|f\| = sup_{x \in X} |f(x)|$.

Definition 3.4 (*Hilbert-Schmidt norm*). $\|\mathcal{O}\|_{\mathcal{HS}} \equiv \|\mathcal{O}\|_2 = \sqrt{\sum_n \|\mathcal{O}|n\rangle\|^2}$, *where* $|n\rangle$ *is an orthonormal basis on* $\mathcal{H}$. *Another way to write this norm is* $\|\mathcal{O}\|_{\mathcal{HS}} = \sqrt{Tr|\mathcal{O}|^2}$.

Definition 3.5. *The trace-norm can be defined as* $\|\star\|_1 : X \longmapsto \|X\|_1 \equiv Tr(|X|) = Tr(\sqrt{X^\dagger X}) = \sum_i x_i$, *where the* x_i *are the singular values of* X.

Definition 3.6. *A Hilbert-Schmidt operator is a bounded operator* $\mathcal{O}$, *with finite Hilbert-Schmidt norm* ($\|\star\|_2$), *on a separable*[3] *Hilbert space* $\mathcal{H}$.

The trace-norm ($\|X\|_1$), the Hilbert–Schmidt norm ($\|X\|_2$) and the largest singular value, defined as $\|X\| \equiv \max_{1 \leq i \leq n}\{x_i\}$, where the x_i are the singular values of X, are all equivalent on finite-dimensional $\mathcal{H}$ [12]. The inequalities for these norms are [12]:

$$\begin{aligned}
\|X\| &\leq \|X\|_1, \\
\|X\| &\leq \|X\|_2, \\
\|X\|_2 &\leq \|X\|_1, \\
\|X\|_1 &\leq n\|X\|, \\
\|X\|_2 &\leq \sqrt{n}\|X\|, \\
\|X\|_2 &\leq \sqrt{n}\|X\|_1.
\end{aligned}$$

Then, the uniform, the trace and the Hilbert-Schmidt norms are all equivalent on finite-dimensional $\mathcal{H}$ and thus define equivalent topologies with the same converging sequences [12].

[2] These two properties are needed if we want to preserve our intuitive concept of distance.
[3] A separable Hilbert space is a Hilbert space such as there exists enumerable basis.

3.1.3 Vector Spaces and Inner Products

For any *ket vector* $|x\rangle \in \mathcal{V}$, there exists an isomorphic space $\mathcal{V}^*$ called the dual space of $\mathcal{V}$. The elements of the complex dual space are the linear functionals $f : \mathcal{C} \longmapsto \mathcal{C}$. We can associate a linear functional "make the inner product with $|x\rangle$", to every ket vector $|x\rangle$. Then we can define a *bra vector* ($\langle x| \in \mathcal{V}^*$) such that the functional $f : \mathcal{C} \longmapsto \mathcal{C}$ will be $f : |x\rangle \longmapsto \langle x|y\rangle$ [20]. Let $\mathcal{C}$ be a complex vector space equipped with an inner product. An *inner product* in $\mathcal{C}$ is a function $\langle\star|\star\rangle : \mathcal{C} \times \mathcal{C} \longmapsto \mathcal{C}$. If $|x\rangle$, $|y\rangle \in \mathcal{V}$ and $\lambda \in \mathcal{C}$, the inner product has the following properties [19]:

1. $(\langle x| + \langle x'|)|y\rangle = \langle x|y\rangle + \langle x'|y\rangle$,
2. $\langle \lambda \cdot x|y\rangle = \lambda^* \langle x|y\rangle$, but $\langle x|\lambda \cdot y\rangle = \lambda \langle x|y\rangle$,
3. $\langle x|y\rangle = \langle y|x\rangle^*$,
4. $\langle x|x\rangle = 0$ *iff* $|x\rangle = 0$.

We can define a norm from the inner product, *i.e.*, $\|x\| \equiv \sqrt{\langle x|x\rangle}$. The two first properties of the norm function are easily satisfied for this norm. We also have the *Cauchy–Schwarz inequality*: $|\langle x|y\rangle| \leq \|x\|\|y\|$. [19].

Definition 3.7. *If A and B are two Hilbert–Schmidt operators, then the Hilbert–Schmidt inner product can be defined as $\langle A|B\rangle_{\mathcal{HS}} \equiv Tr(A^\dagger B) = \sum_n \langle n|A^\dagger B|n\rangle$.*

3.2 The Algebra of the Classical and Quantum Observables

This section[4] is an attempt in order to enlighten another point of view toward the classical and quantum mechanics formalism. Instead of defining states, we define the observables and the set of rules that they must obey. In this "operational" theory, the states are functionals of the space in which we can attribute probabilities [23]. The expected values of the observables define the state of the system. Then in this section we try to understand in a simple way the algebra of the quantum and classical observables.

Both quantum and classical systems are well described by the structure of the C^*-algebra[5] of the observables. From an operational point of view, a *physical system* is fully described by its physical properties, *i.e.*, by the set Θ of the *observables* [21]. For any observable $f \in \Theta$ and $\lambda \in \mathcal{R}$, one can define the observable λf, which is a re-scaling of the apparatus by λ [21].

A state ω of a physical system is characterized by the results of the measurements of the observables in the sense that the *average* over the results of measurements of an observable f, when the system is in a state ω, defines the *expectation* value $\omega(f)$. Thus the state is completely characterized by *all* its expectations $\omega(f)$, when f varies over the set Θ. Then we can say that ω is a real functional on Θ [21]. This operational characterization of the states in terms of its expectations of the observables requires that two states yielding

[4] I would like to thank M. Terra-Cunha, [20, 21] and also Bárbara et al. [22] for the idea followed in this next section.

[5] Algebra is the branch of mathematics that studies the rules of operations and relations.

the same expectations must be identical, *i.e.,* if $\omega_1(f) = \omega_2(f)$ for all $f \in \Theta$, then $\omega_1 = \omega_2$, and on the other hand if f and g have the same expectations $\omega(f) = \omega(g)$, for all states ω, then we cannot distinguish these observables, that is, $f = g$ [21].

3.2.1 Some C^*-Algebra Properties

A Banach-algebra ($\mathcal{A}$) is a vector space and it is also a Banach space [22, 24]. A Banach space is also a metric space, *i.e.,* a space where there exists the concept of *distance*.

Definition 3.8. *Let $f, g \in \mathcal{A}$. An involution is a map $* : \mathcal{A} \longmapsto \mathcal{A}$ and the image of and element f by the involution map is $f \longmapsto f^*$. The involution has the following properties:*

1. $(f+g)^* = f^* + g^*$.
2. $(fg)^* = g^* f^*$.
3. For every $\lambda \in \mathcal{C}$, $\quad (\lambda f)^* = \bar{\lambda} f^*$.
4. For every $f \in \mathcal{A}$, $\quad (f^*)^* = f$.

A C^*-algebra is an algebra over the field of the complex numbers ($\mathcal{C}$), where the involution map can be defined. The space of bounded linear transformations of a Hilbert space $\mathcal{H}$ is a C^*-algebra. Suppose that $f, g \in C^*$ then

1. $\|fg\| \leq \|f\|\|g\|$.
2. $\|f^*\| = \|f\|$.
3. $\|f\|^2 = \|ff^*\| = \|f^*f\| = \|f\|\|f^*\|$.
4. There exists 1 such as $1f = f1 = f$ for every f and $\|1\| = 1$.

3.2.2 The Algebra of the Classical Observables

In a rigorous description of statistical mechanics [23], the *observables* are *functions* $f \in \mathcal{A}$ in the phase-space and the *states* are *functionals* $\omega(\star)$ defined over the observables and we can assign probabilities with these functionals [25].

The observables associated to a classical system generate an abelian[6] algebra ($\mathcal{A}$) of real, or more generally complex continuous functions f on the phase space [21].

This algebra obeys the C^* conditions: the identity is given by the function $f = 1$, the product is given by $(fg)(x) = f(x)g(x)$, the involution map is given by the complex conjugation $f(x)^* = \bar{f}(x)$. Finally, we also have the following property $\|f^*f\| = \|f\|^2$. We can also assign a sup-norm (Def. 3.3) for each $f \in \mathcal{A}$ and it can be shown (see [21]) that this algebra is a Banach space with

[6] An abelian algebra is an algebra in which the multiplication operation always obeys the commutative law.

respect to the norm topology (the sup-norm). Summarizing, the algebra of the classical observables follows the following rules:

1. Observables are functions on the phase-state: $f \in \mathcal{A}$.
2. Observables form a commutative (abelian) C^*-algebra, and they obey the C^* conditions, see Subsection 3.2.1.
3. States are linear functionals $\omega : \mathcal{A} \to \mathcal{C}$, with $f \longmapsto \omega(f)$.

3.2.3 States of a Commutative Algebra

A state of a system is characterized by the measurements of the observables in that state. The expectation value of the observable f on the state ω is given by: $\omega(f) = \lim_{n\to\infty}\langle f\rangle_n$. Since the expectation $\omega(f)$ has the interpretation of the average of the measurements of f in the given state ω, it follows that the expectations are linear functionals, *i.e.*, $\omega(\lambda_1 f + \lambda_2 g) = \lambda_1\omega(f) + \lambda_2\omega(g)$.

Definition 3.9 (States). *A state is linear functional $\omega : \mathcal{A} \longmapsto \mathcal{C}$, which has the following properties* [22]*:*

1. $\omega(ff^*) \geq 0$, *(positivity).*
2. $\omega(1) = 1$, *(normalization).*

The states must be normalized, this property is trivial since we can always change $\omega \to \omega(1)^{-1}\omega$ [21]. The linear functional called expectation must be positive in order to preserve the Cauchy-Schwarz inequality [21]:

Theorem 3.10. *The positivity of the functional ω, i.e., $\omega(ff^*) \geq 0$, for all f, implies the validity of the Cauchy-Schwarz inequality.*

Proof: Let $f = x + \lambda y$ where $x, y \in \mathcal{A}$ and $\lambda \in \mathcal{C}$. We need to show that if the condition 1. of Def. 3.9 holds, *i.e.*, if $\omega(ff^*) \geq 0$, then the Cauchy-Schwarz is valid. Of course $ff^* \geq 0$. So $\omega(ff^*) \geq 0$ implies that $|\langle x|y\rangle| \leq \|x\|\|y\|$, that is, $|\omega(\langle x|y\rangle)|^2 \leq \omega(xx^*)\omega(yy^*)$. Thus $f \in \mathcal{C}$, $\sqrt{ff^*} = \|f\| \geq 0$, $\forall\lambda \in \mathcal{C}$,

$$\begin{gathered}
\omega(ff^*) = \omega(\|x + \lambda y\|^2) \geq 0, \\
\omega(\|x\|^2) + \omega(|\lambda|^2\|y\|^2) + \omega(\langle x|\lambda y\rangle) + \omega(\langle \lambda y|x\rangle) \geq 0, \\
\omega(\|x\|^2) + |\lambda|^2\omega(\|y\|^2) + \lambda\omega(\langle x|y\rangle) + \lambda^*\omega(\langle y|x\rangle) \geq 0, \\
\begin{pmatrix} 1 & \lambda^* \end{pmatrix} \begin{pmatrix} \omega(\|x\|^2) & \omega(\langle x|y\rangle) \\ \omega(\langle y|x\rangle) & \omega(\|y\|^2) \end{pmatrix} \begin{pmatrix} 1 \\ \lambda \end{pmatrix} \geq 0, \\
\begin{pmatrix} 1 & \lambda^* \end{pmatrix} M \begin{pmatrix} 1 \\ \lambda \end{pmatrix} \geq 0,
\end{gathered}$$

As the matrix M is a positive matrix, ($M \geq 0$), *i.e.*, $\langle\psi|M|\psi\rangle \geq 0$, $\forall|\psi\rangle \in \mathcal{H}$, thus its determinant must be positive. $det(M) \geq 0$ implies that $|\omega(\langle x|y\rangle)|^2 \leq \omega(\|x\|^2)\omega(\|y\|^2)$, or $|\omega(\langle x|y\rangle)|^2 \leq \omega(xx^*)\omega(yy^*)$ [26]. A more beautiful proof for Theorem 3.10 can be done with the help of the positive maps and the Jamiołkowski isomorphism (see [27]). For a sketch of the proof see Chapter 7.

3.2.4 The Algebra of the Quantum Observables

The Quantum Mechanical systems are ruled by C^* non-commutative algebra of the observables. Classical systems are described by points in the phase-space and by a commutative algebra over $\mathcal{R}$ and its trajectories in the phase space are given by deterministic Hamiltonians [21].

However, in finite dimensional quantum mechanics, the matrices form another vector space: the non-commutative $M_n(\mathcal{C})$. This set is formed by the $n \times n$ matrices with complex coefficients. This vector space has a natural inner product, (Def. 3.7), which turns it into a Banach space [22]. A natural norm for this space is the usual operator norm:

Definition 3.11 (Usual Operator Norm)**.** $\|\mathcal{O}\| = \sup_{f:|f|=1} |\mathcal{O}(f)|$.

Hence, we can created a similar structure to the classical case and think about the *quantum observables* as linear functionals that form an *algebra*, (the C^* non-commutative algebra). They are the Hermitian matrices $\mathcal{O} \in M_n(\mathcal{C})$, and the *states* are defined by positive functionals (the density matrices) over the observables and we can also relate these positive functionals with probabilities [23, 25, 28].

If we define the involution map as $A^* = A^\dagger$, *i.e.*, $(A)^*_{ij} = \bar{a}_{ji}$, we will have a non-commutative C^*-algebra, (see for example [22]). The functional $\omega(\mathcal{O}^\dagger\mathcal{O}) \geq 0$ implies the positivity of $\mathcal{O}^*\mathcal{O} = \mathcal{O}^\dagger\mathcal{O} \geq 0$. Summarizing, the algebra of the quantum observables follows:

1. Observables are operators $\mathcal{O} \in M_n(\mathcal{C})$.
2. Observables form a non-commutative C^*-algebra, $(M_n(\mathcal{C}))$, and they obey the C^* conditions, see Subsection 3.2.1.
3. States are linear functionals $\omega : M_n(\mathcal{C}) \to \mathcal{C}$, with $\mathcal{O} \longmapsto \omega(\mathcal{O})$.

3.2.5 States of the Non-Commutative Algebras

Definition 3.12 (States)**.** *A state is linear functional* $\omega : M_n(\mathcal{C}) \longmapsto \mathcal{C}$*, with* $\mathcal{O} \longmapsto \omega(\mathcal{O})$*, which has the following properties* [22]*:*

1. $\omega(\mathcal{O}\mathcal{O}^\dagger) \geq 0$*, (positivity).*
2. $\omega(\mathbb{I}) = 1$*, (normalization).*

A representation of a linear functional ω is given by the trace one positive Hermitian matrices $\rho \in M_n(\mathcal{C})$. If we define the linear functional expectation $\langle\mathcal{O}\rangle_\rho = \langle\rho|\mathcal{O}\rangle = Tr(\rho\mathcal{O})$, we will assign the linear functional $\omega(\star) = Tr(\rho\star)$ with a quantum state, *i.e.*, with its density matrix. Of course $\omega(\mathcal{O}) = Tr(\rho\mathcal{O})$.

3.3 Convex Sets

In a geometric description of Quantum Mechanics, it is natural that some questions could arise on the restrictions required in those sets. As occurs in statistical mechanics, we intend to define the convex mixtures of elements of the set.

The set of the density matrices for pure and mixed states is a convex set. A *convex set* is a set such that every *convex mixture* of its points belongs to the set. However, this definition needs to be improved. We can always define convex mixtures in a geometrical sense, by first defining a straight line between any two points belonging to a convex set.

In this geometric point of view, a convex mixture of two quantum states is defined by a point (another quantum state) in the straight line between these two states.

Definition 3.13 (Straight Line)**.** *A straight line between two points x and y is a set of points which $z = ax + by$ and $a + b = 1$.*

Definition 3.14 (Convex Set)**.** *A subset C of the Euclidean space $\mathcal{E}^n$ is a convex set if for all pairs $x, y \in C$, the convex mixture defined by $z = ax + by$ with $a + b = 1$ and $a \geq 0$, $b \geq 0$ also belongs to the set, i.e., if $z \in C$.*

As an example of convex mixture, suppose for instance we have m_x kg of a coffee x and m_y kg of a coffee y, with fixed prices p_x and p_y respectively. We can produce another coffee z as a mixture of these two products. Note that subtraction of mass is forbidden here. The price of the coffee z is $p_z = \frac{m_x p_x + m_y p_y}{m_x + m_y}$. This is evidently a convex mixture. Suppose that $p_x < p_y$, then of course we always have $p_x \leq p_z \leq p_y$. And if we define $a \equiv m_x/(m_x + m_y)$ and $b \equiv m_y/(m_x + m_y)$, then $p_z = ap_x + bp_y$ with $0 \leq a, b \leq 1$ and $a + b = 1$, this equation is identical to the definition of a convex set given by Def. 3.14.

Definition 3.15 (Pure Point)**.** *A pure point of a convex set C is a point which cannot be obtained by any mixture of points $x, y \in C$. The non-pure points are called mixed.*

Definition 3.16 (Quantum Pure State)**.** *A quantum pure state is a complex vector $|\psi\rangle \in \mathcal{H}$, where $\mathcal{H}$ is n complex-dimensional Hilbert space.*

Asher Peres in [15, 29] says that a quantum pure state is a state for which there exists answers with probability $p = 1$ for a certain number of questions.

Consider a physical system described by a pure vector $|\psi\rangle$. Then there exists a basis, namely $\{|i\rangle, i = 1, ..., d\}$ that expands the subspace of $\mathcal{H}$ where the vector $|\psi\rangle$ lives. This expansion can be written as $|\psi\rangle = \sum_{i=1}^{d} c_i|i\rangle$, with $\sum_{i=1}^{d} |c_i|^2 = 1$.

A physical observable A is a Hermitian operator with matrix elements given by $A_{ij} = \langle i|A|j\rangle$ in this basis. The expected value of this observable in the state $|\psi\rangle$ is given by $\langle A\rangle_\psi = \langle\psi|A|\psi\rangle$. The complete set of the postulates of Quantum Mechanics for the *braket* formalism can be found in any intermediate book as [2] or [30].

Definition 3.17. *A convex hull is the smallest convex set that contains the set. The convex hull of a finite set of points is called a convex polytope.*

Definition 3.18. *A p-simplex is a set of $p+1$ points not confined to any $(p-1)$-dimensional subspace* [31].

Theorem 3.19 (Minkowski)**.** *Any convex body is the convex hull of its pure points* [31].

Theorem 3.20 (Carathéodory)**.** *The rank of a point in a convex hull is the minimum number r necessary to express this point as a convex combination of rank one pure states.*

$$x = \sum_{i=1}^{r\leq n+1} c_i x_i, \text{ with } c_i \geq 0 \text{ and } \sum_{i=1}^{r} c_i = 1.$$

Any density matrix of rank n can not be written as a convex combination of less than n pure states (projectors). The set of the diagonal density matrices, which are diagonal in one chosen basis form a $(n-1)$-dimensional convex set known as *simplex of eigenvalues* that is a polytope centered by the trace one maximally mixed state $\mathbb{I}$. Then, our definition of rank must be improved. When we are dealing with mixture of pure points in a simplex, the rank of the point is given by Theorem 3.20. However these points consist of diagonalizable matrices, then we will need to define another concept of rank, and it will be done later.

Theorem 3.21 (Hahn-Banach separation theorem)**.** *Given a closed convex body $\mathcal{C}$ and a point z in the exterior of this body, there exists a linear functional W that takes non-negative values for all points of this convex body ($W(x) \geq 0,\ \forall x \in \mathcal{C}$), while $W(z) < 0$.*

Proof: The proof is out of the scope of this text. For our purposes here it is sufficient to discuss the Fig. 3.21. This figure shows a closed convex set ($\mathcal{C}$) and a point z located outside the convex body. It is obvious that there are linear functionals or hyperplanes ($W_1, W_2, ...$) which separate the point z of the set $\mathcal{C}$. A simple and beautiful proof can be seen in [32], (look for Theorem 1.11).

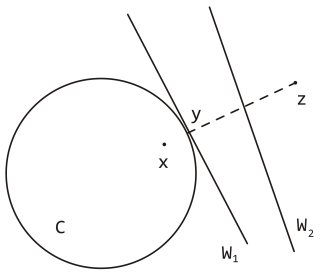

Figure 3.1: The Hanh-Bannach Hyperplane Separation Theorem

3.4 The Density Matrix Formalism

The *braket* formalism has some disadvantages. It is fine when we are dealing just with pure states, but if we do not possess the complete description or complete information of a system, we claim for the concept of *probabilities*. In Chapter 2, we described pure states of polarization using the *Jones vector formalism*. We showed that all the information of a pure state of polarization is contained in a *Jones vector*. But in order to describe partially polarized states, *i.e.*, when $0 \leq P < 1$, we needed to invoke the *coherence matrix formalism*. We showed that both totally and partially polarized states can be described by this latter formalism. Also in Chapter 2, we presented some of the clear analogies between the coherence matrix J and the density matrix for one qubit ρ and its representation in the *Bloch-Poincaré* ball and sphere. Now we are going to build a formalism to describe, in a more appropriate way, the pure and mixed quantum states for more than one qubit.

Suppose, for example, that we have a source (an oven in a Stern-Gerlach experiment) which sends atoms of Silver described by vectors $|n\rangle$ with energy equal to E_n, and with probability proportional to $e^{\frac{-E_n}{KT}}$, *i.e.*, the state $|1\rangle$ has energy E_1 and with probability p_1, so forth[7]. We could assign each state with its probability by constructing the following set $\{p_i, |i\rangle\}$, apply the postulates of Quantum Mechanics for all i, and, later developing some statistics with these data. But we have a much more appropriate formalism to work this *quantum ensemble*. As discussed later the formalism that can operate with pure states vectors and with statistical mixtures is the *density matrix formalism*. Firstly we need to define precisely the density matrix for a quantum pure state (Def. 3.22) and the density matrix for quantum convex statistical mixtures, *i.e.*, for quantum mixed states[8] (Def. 3.23).

Definition 3.22 (Density Matrix for Pure states)**.** *A density matrix for a quantum pure state $|\psi\rangle$ is usually called ρ_ψ or simply ρ and it is given by the following rule: $\rho_\psi = |\psi\rangle\langle\psi|$.*

Note that the density operator is a projector as defined by its properties in the Eq. 2.8 and it is obviously a Hermitian operator ($\rho^\dagger = \rho$). It is a positive semidefinite operator as it is shown in Section 3.4.1. This operator does not depend on the global phase of the vector $|\psi\rangle$, because if we define another vector $|\psi'\rangle \equiv e^{i\theta}|\psi\rangle$, with $\theta \in \mathcal{R}$, then their density operators are the same, *i.e.*, $\rho_{\psi'} = \rho_\psi$, since $\langle\psi'| = e^{-i\theta}\langle\psi|$.

Suppose we have a Hermitian observable A. As seen before, the expected value of this observable is $\langle A\rangle = \langle\psi|A|\psi\rangle$. In the density operator formalism, this expected value can be written as $\langle A\rangle = Tr(A|\psi\rangle\langle\psi|) = Tr(A\rho_\psi)$. The proof is easy, just insert a closure relation $\sum_i |i\rangle\langle i| = \mathbb{I}$ inside the sandwich $\langle A\rangle = \sum_i \langle\psi|A|i\rangle\langle i|\psi\rangle = \sum_i \langle i|\psi\rangle\langle\psi|A|i\rangle = Tr(\rho_\psi A)$.

[7] Of course the states here are not all orthogonal, since in a two dimensional space, given a vector we can find just another orthogonal vector.

[8] You can find another good approach for this discussion in [20].

Definition 3.23 (Mixed Quantum State). *A mixed quantum state is a convex mixture of pure states* $|\psi_i\rangle\langle\psi_i|$, $\rho = \sum_{i=1}^{d} p_i |\psi_i\rangle\langle\psi_i|$, *with* $\sum_{i=1}^{d} p_i = 1$ *and* $p_i \geq 0 \quad \forall i$.

The statistical mixture of pure states $\{|1\rangle, |2\rangle, ..., |d\rangle\}$ with probabilities $\{p_1, p_2, ..., p_d\}$ can be written exactly as discussed before, just re-label the operators $|i\rangle\langle i|$ as $|\psi_i\rangle\langle\psi_i|$.

Theorem 3.24. *The pure states are the pure points of the convex set of the density matrices, i.e., this set is a convex set whose extreme points are the pure states.*

Proof: It is obvious that all convex combinations of density matrices are positive and trace one, then the set of the density matrices is obviously a convex set. We need to show that the extreme points of this set, (*the pure points*), are the pure states.

Suppose that a pure state $|\psi\rangle\langle\psi| \equiv P_\psi$ can be written in a convex combination of two other density matrices ρ and σ. So $P_\psi = \lambda\rho + (1-\lambda)\sigma$, with $0 \leq \lambda \leq 1$. Then if we multiply both sides by P_ψ, we will have $P_\psi P_\psi = \lambda\rho P_\psi + (1-\lambda)\sigma P_\psi$. Tracing both sides, and using $P_\psi^2 = P_\psi$, $Tr(P_\psi) = 1$, and $Tr(\rho P_\psi) \equiv \langle\rho|P_\psi\rangle$, we have $1 = \lambda\langle\rho|P_\psi\rangle + (1-\lambda)\langle\sigma|P_\psi\rangle$, by using the Cauchy-Schwarz inequality, $0 \leq |\langle\rho|P_\psi\rangle|^2 \leq \langle\rho|\rho\rangle\langle P_\psi|P_\psi\rangle \leq 1$, the two last equalities in the middle holds *iff* $\rho = cP_\psi$, and $\sigma = dP_\psi$, where $c, d \in \mathcal{C}$, and $|c|, |d| = 1$. By using the trace conditions, we find that $c = d = 1$. Finally, $\langle\rho|P_\psi\rangle = \langle\sigma|P_\psi\rangle = 1$, and this shows that it is impossible to make convex combinations of two different states to form a pure state.

Of course we could impose, from the beginning, the purity of ρ and σ, because a non-trivial convex mixture of interior points can not create points in the boundary of the convex set. Then we can construct another proof for Theorem 3.24 writing a state P_ψ in its spectral decomposition. Since P_ψ is a hermitian operator, it can be diagonalized. Then $P_\psi = \sum_{i=1}^{n} \lambda_i |i\rangle\langle i|$, with $0 \leq \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$. If we multiply both sides by P_ψ, we will have $P_\psi^2 = \sum_{i,i'}^{n} \lambda_i \lambda_{i'} |i\rangle\langle i||i'\rangle\langle i'|$. We know that eigenvectors related to different eigenvalues in a hermitian operator are orthogonal, then $P_\psi^2 = \sum_{i,i'}^{n} \lambda_i^2 |i\rangle\langle i|$. Finally, using the property: $P_\psi^2 = P_\psi$, we will have $\lambda_i^2 = \lambda_i$, that is, $\lambda_i = 0$ or 1. Using the fact that $Tr(P_\psi) = 1$, we see that there exists only one $\lambda_i = 1$ and all the other eigenvalues are equal to 0. □

3.4.1 The Postulates of Quantum Mechanics

The postulates in this section are a *resumé* of the postulates and they can be found in [10].

1. **States** – We elevate the states to the category of the operators and they are represented by their density matrices (trace one Hermitian positive operators) in a space called Hilbert–Schmidt space ($\mathcal{H}_{\mathcal{HS}}$). These operators can be used in order to describe probability distributions of quantum states.

- $Tr(\rho) = 1$. Suppose that $\rho = \sum_i^d p_i|\psi_i\rangle\langle\psi_i|$ and $\sum_{i=1}^d p_i = 1$. (If ρ is a pure state, just change the summation over i by a unique term $|i\rangle\langle i|$, because $p_i = 1$ for some i). Then we will have $Tr(\rho) = \sum_{i=1}^d p_i Tr(|\psi_i\rangle\langle\psi_i|) = \sum_{i=1}^d p_i = 1$. Remember that the projectors have rank and trace equal to one.
- $\rho \geq 0$. This means that all eigenvalues of ρ are non-negative, or in other words, ρ is positive semidefinite. Suppose $|\phi\rangle$ is an arbitrary vector in Hilbert space $\mathcal{H}$. Its "expected value" can be calculated as $\langle\phi|\rho|\phi\rangle = \sum_{i=1}^d p_i\langle\phi|\psi_i\rangle\langle\psi_i|\phi\rangle = \sum_{i=1}^d p_i|\langle\phi|\psi_i\rangle|^2 \geq 0$. As this is true for any $|\phi\rangle \in \mathcal{H}$, we can conclude that $\rho \geq 0$.

2. **Observables** – The observables are Hermitian operators $E_i \in \mathcal{H}_{\mathcal{HS}}$.

3. **Measurement** – A measurement operator is a positive operator E_i that can be written as $E_i = M_i^\dagger M_i$ for some operator M_i. This operator E_i is an element of the positive operators set $\{E_1, ..., E_n\}$ and this set obeys the closure relation $\sum_{i=1}^n E_i = \mathbb{I}$.

   When a measurement is made in a quantum state ρ, the *ith*-outcome appears with probability $p_i = Tr(\rho M_i^\dagger M_i)$. Suppose that we perform a measurement described by the *braket* formalism. We know that the probability of obtaining an outcome m, given the state is represented by $|\psi_i\rangle$, is described by $p(m|i) = \langle\psi_i|M_m^\dagger M_m|\psi_i\rangle$. This expression can be written in the density matrix formalism as $Tr(M_m^\dagger M_m \rho)$. By using the probabilities law we have $p(m) = \sum_{i=1}^d p(m|i)p_i = \sum_{i=1}^d p_i Tr(M_m^\dagger M_m|\psi_i\rangle\langle\psi_i|) = Tr(M_m^\dagger M_m\rho)$.

   How could we write the density operator after performing a measurement and obtaining the *mth*-outcome? If the initial state is given by a pure vector $|\psi_i\rangle$, then the final state after obtaining the *mth*-outcome will be $\left|\psi_i^m\right\rangle = \frac{M_m|\psi_i\rangle}{\sqrt{\langle\psi_i|M_m^\dagger M_m|\psi_i\rangle}}$. By the Bayes rule, (see Section 4.3.1), $p(i|m) = \frac{p(m,i)}{p(m)} = \frac{p(m|i)p(i)}{p(m)}$. After the measurement which yields the result m, we have an ensemble of states $\left|\psi_i^m\right\rangle$ with probabilities $p(i|m)$. Then the density operator is given by:

$$\rho_m = \sum_i p(i|m)|\psi_i^m\rangle\langle\psi_i^m|,$$

$$\rho_m = \sum_i p(i|m)\frac{M_m|\psi_i\rangle}{\sqrt{\langle\psi_i|M_m^\dagger M_m|\psi_i\rangle}}\frac{\langle\psi_i|M_m^\dagger}{\sqrt{\langle\psi_i|M_m^\dagger M_m|\psi_i\rangle}},$$

$$\rho_m = \sum_{i=1}^d p_i\frac{M_m|\psi_i\rangle}{\sqrt{\langle\psi_i|M_m^\dagger M_m|\psi_i\rangle}}\frac{\langle\psi_i|M_m^\dagger}{\sqrt{\langle\psi_i|M_m^\dagger M_m|\psi_i\rangle}} = \frac{M_m\rho M_m^\dagger}{Tr(\rho M_m^\dagger M_m)}.$$

4. **Time Evolution** – The time evolution of quantum systems is given by a *CP*-map [9], *i.e.*, a Completely Positive Map. Then the state $\rho(t)$ is

[9] For a complete review of the *CP*-maps, see [27].

given by $\rho(t) = \sum_i K_i(t)\rho(0)K_i^\dagger(t)$, where K are the Kraus operators. The *Schrödinger* evolution is a special case of this evolution when the K are unitary matrices U, and the state in time t, *i.e.*, $\rho(t)$ is given by $\rho_s(t) = U(t)\rho(0)U^\dagger(t)$. If the environment is not perfectly well known, then $\rho_s(t) = \sum_i p_i U_i(t)\rho(0)U_i^\dagger(t)$.

3.5 The Space Of More Than One Qubit

3.5.1 The Tensor Product

In order to describe more than one qubit, we need the concept of the *tensor product*. When we create two photons in a laboratory, we need to know how to write the global quantum state of these two photons, *i.e.*, we need to know how to compose physical systems (for example, two systems of two levels) and we also need to define and to perform *local operations* in one part of these systems.

Definition 3.25. *Let there be $A_{(m\times n)}$ and $B_{(p\times q)}$ two matrices. Then the tensor product of this two matrices, i.e., the matrix $A \otimes B$ can be defined by its matrix elements: $(A \otimes B)_{(m\cdot p)\times(n\cdot q)}$* [13].

Suppose we have a quantum harmonic oscillator A described in a Hilbert $\mathcal{L}_A^2$ space where the Hermite functions $\phi_n(x), \forall\ n \in \mathcal{N}$ form a complete basis. Let us consider another quantum harmonic oscillator B oscillating in y-axis. Then we need another Hilbert space $\mathcal{L}_B^2$ and another set of Hermite functions $\psi_m(y), \forall\ m \in \mathcal{N}$ in order to describe this other degree of freedom. We also need the concept of tensor product in order to construct a function which describes these two degrees of freedom, *i.e.*, the set $\{\phi_n(x)\psi_m(y),\ \forall\, n,\ m \in \mathcal{N}\}$ forms a basis for the state of space of this two dimensional quantum harmonic oscillator. Any squared-integrable function $f(x, y)$ can be expanded in such basis. In Dirac notation, this basis can be expressed as $\{|\phi_n\rangle \otimes |\psi_m\rangle,\ \forall\, n,\ m \in \mathcal{N}\}$.

Let us return to the spaces $\mathcal{H}_\mathcal{A}$ and $\mathcal{H}_\mathcal{B}$. Then we can define a third Hilbert space $\mathcal{H}_{\mathcal{AB}}$ and a bilinear map T such as $\mathcal{H}_{\mathcal{AB}} = \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$ and all the following rules apply (adapted from [30]).

1. $T(\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{B})$ generates $\mathcal{H}_{\mathcal{AB}}$, *i.e.*, any vector $|\psi\rangle \in \mathcal{H}_{\mathcal{AB}}$ can be written as a sum of the form $T(|u\rangle, |v\rangle)$, that is a sum of vectors $|u\rangle \in \mathcal{H}_\mathcal{A}$ and $|v\rangle \in \mathcal{H}_\mathcal{B}$.
2. Consider a basis $\{|u_i\rangle\}$ of $\mathcal{H}_\mathcal{A}$ and $\{|v_j\rangle\}$ of $\mathcal{H}_\mathcal{B}$. Then the set $\{T(|u_i\rangle, |v_j\rangle)\}$ forms a basis of $\mathcal{H}_{\mathcal{AB}}$.
3. $T(\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{B}) \equiv \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$, and $T(|u\rangle, |v\rangle) \equiv |u\rangle \otimes |v\rangle$.

Of course all definitions and properties can be extended for more than two parts. If each *ith*-degree of freedom is described by a particular Hilbert space $\mathcal{H}_i$, then any system involving n degrees of freedom can be described as $\mathcal{H}_n = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_n$. We can enumerate some properties of the tensor product (Also adapted from [30]):

- $dim(\mathcal{H}_{\mathcal{AB}}) = dim(\mathcal{H}_{\mathcal{A}})dim(\mathcal{H}_{\mathcal{B}})$.
- Let $|\psi_A\rangle = |u\rangle \otimes |v\rangle \in \mathcal{H}_{\mathcal{AB}}$ and $|\psi_B\rangle = |u'\rangle \otimes |v'\rangle \in \mathcal{H}_{\mathcal{AB}}$ with $|u\rangle, |u'\rangle \in \mathcal{H}_{\mathcal{A}}$ and $|v\rangle, |v'\rangle \in \mathcal{H}_{\mathcal{B}}$. Then the inner product $\langle\psi_A|\psi_B\rangle$ of this two vectors can be defined in the space $\mathcal{H}_{\mathcal{AB}}$ as $\langle\psi_A|\psi_B\rangle = \langle u|u'\rangle_A \otimes \langle v|v'\rangle_B = \langle u|u'\rangle_A \cdot \langle v|v'\rangle_B$.
- If an operator A acts in the space $\mathcal{H}_{\mathcal{A}}$ and B acts in the space $\mathcal{H}_{\mathcal{B}}$, then the operator $A \otimes B$ acts in $\mathcal{H}_{\mathcal{AB}} = \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$. Any operator that can be written in this form is a *local operator*. Any laboratory operation described by a tensor product is a *local operation*.
- It is possible to act only in $\mathcal{H}_{\mathcal{A}}$ or only in $\mathcal{H}_{\mathcal{B}}$ of the product space $\mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}} = \mathcal{H}_{\mathcal{AB}}$. Suppose that A acts in $\mathcal{H}_{\mathcal{A}}$ and B acts in $\mathcal{H}_{\mathcal{B}}$. Let us define two new operators $A \otimes \mathbb{I}_B$ and $\mathbb{I}_A \otimes B$ acting on the product space $\mathcal{H}_{\mathcal{AB}}$. They are called extensions of A and B respectively.

If two systems do not have any correlation and if we use the *tensor product* and the *Born's rule*, we will get our classical concept of *statistical independence* for composite systems. Suppose that a system is described by $|\psi\rangle$. If we measure a physical property A, the only possible results are the eigenvalues of A. Let us suppose that the eigenvalue a is related to the eigenvector $|a\rangle$. Then we will obtain a result a with probability $p(a) = |\langle a|\psi\rangle|^2$. Analogously, if we measure another physical property B in another system which is described by $|\phi\rangle$ we will obtain an outcome b with probability $p(b) = |\langle b|\phi\rangle|^2$. Then $p(a, b)$ can be described by $p(a, b) = p(a)p(b)$, since they are independent. Then it has $p(a, b) = p(a)p(b) = |\langle a|\psi\rangle|^2|\langle b|\phi\rangle|^2 = |\langle a| \otimes \langle b||\psi\rangle \otimes |\phi\rangle|^2$. This suggests that we could consider the system in a state $|\psi\rangle \otimes |\phi\rangle \in \mathcal{H}_{\mathcal{AB}} \equiv \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$, with local measurements given by $A \otimes \mathbb{I}_B$ and $\mathbb{I}_A \otimes B$. It is clear that this rule of composition preserves our intuitive concept of statistical independence when the two systems are not related.

We already showed that we can capture our intuitive concept of statistical independence by using the tensor product properties and the Born's rule. It is enough to suppose that nature constructs her composite systems adopting the tensor product hypothesis. But the opposite was not considered here, *i.e.*, we do not know if the statistical independence implies the tensor product hypothesis. Coming from a philosophical point of view, we could construct composite systems by making other assumptions such as the commutative assumption [33]. Probably nature does not construct composite systems from subsystems, rather, she presents us composite systems which we perceive as made of subsystems [34]. However this is not the point here, we understand that Quantum Mechanics Theory works well with the tensor product assumption, thus for our purposes here, it is quit enough.

3.6 The Hilbert–Schmidt Space

As discussed in Section 3.2.1, our quantum states are density matrices on a vector space that is also a non-commutative *algebra*. Let us define a n complex-dimensional Hilbert space $\mathcal{H}$. Then there exists a *dual* Hilbert space $\mathcal{H}^*$

defined as the space of the linear functionals from $\mathcal{H}$ to the complex numbers set. In finite-dimensional case, these two sets are isomorphic [31]. Let us define the Hilbert–Schmidt space as the n^2 dimensional space of the bounded operators in Hilbert-Schmidt norm and it is defined as $\mathcal{H}_{\mathcal{HS}} \equiv \mathcal{H} \otimes \mathcal{H}^*$. An interpretation for this space seems obvious when we write any operator as $\mathcal{O} = c_{1,1}|1\rangle\langle 1| + c_{1,2}|1\rangle\langle 2| + c_{2,1}|2\rangle\langle 1| + \cdots$. But how many terms does this summation have? If we try to use Theorem 3.20, the rank of a point in a convex set is the minimum number $r \leq n^2$ of pure points that are needed in a convex set to express it in a convex combination. This number r, given by this theorem, is a large upper bound, then we need to find another number r, in order to express in a better way the rank of a density matrix. As every Hermitian matrix can be diagonalized, then the usual definition of *rank* of a matrix coincides with our purposes.

Definition 3.26. *If* $\rho|e_i\rangle = \lambda_i|e_i\rangle$, *with* $\sum_{i=1}^{r}\lambda_i = 1$ *and* $\rho = \sum_{i=1}^{r}\lambda_i|e_i\rangle\langle e_i|$, *then the* $rank(\rho) \equiv r \leq N$.

Remark: The Hilbert-Schmidt norms are not C^* norms. An example of this fact is that $\left\|X^\dagger X\right\|_1 = \sqrt{\sum_{i=1}^{n} x_i^2} \neq \sum_{i=1}^{n} x_i = \|X\|_1^2$, and another example is $\left\|X^\dagger X\right\|_2 = \sqrt{\sum_{i=1}^{n} x_i^4} \neq \sum_{i=1}^{n} x_i^2 = \|X\|_2^2$. Then equipped by the Hilbert-Schmidt norm defined in Def. 3.4 and by the Hilbert-Schmidt inner product defined in Def. 3.7, we can define another Hilbert space of operators acting on $\mathcal{H}$ [31]. This is the Hilbert-Schmidt space formed by the finite dimensional bounded operators. This inner product gives raise to an Euclidean distance, the Hilbert-Schmidt distance defined by:

Definition 3.27 (The Hilbert-Schmidt Distance ($D^2_{\mathcal{HS}}$))**.** *The Hilbert-Schmidt Distance is a kind of Euclidean distance* ($D_2^2(A, B)$) *and is defined as:*
$D^2_{\mathcal{HS}} = \frac{1}{2}Tr[(A-B)(A^\dagger - B^\dagger)] = D_2^2(A, B)$.

3.7 A Simple Geometric Interpretation For The Entanglement

The set of the quantum states (D) is a convex set, indeed, convex mixtures of density operators are density operators. Then we can prepare convex mixtures of pure states to generate mixed states as we mentioned in Def. 3.23. We also know by Theorem 3.24 that we can not obtain pure states by performing a mixture of other states in a convex mixture. We now define another convex set, the set of separables (S), and a state is called entangled if it belongs to the set $D \setminus S$.

Definition 3.28. *A pure bipartite and separable quantum state is the tensor product of two quantum pure states.* $\rho = \rho^A \otimes \rho^B$.

We can write[10] these bipartites *pure and separable* quantum states given in Def.3.28 as $\rho = |\psi_A\rangle\langle\psi_A| \otimes |\phi_B\rangle\langle\phi_B|$ or $\rho = |\psi_A\phi_B\rangle\langle\psi_A\phi_B|$. All convex combinations of these *separable* states define a convex set (S) called the set of *separable* states.

Definition 3.29. *A mixed and separable bipartite quantum state is defined as a convex combination of pure separable states* $\rho = \sum_i \lambda_i \rho_i^A \otimes \rho_i^B$ *with* $\sum_i \lambda_i = 1$ *and* $\lambda_i \geq 0 \, \forall i$.

Of course if $\lambda_i = 1$, for some i the state in Def. 3.29 becomes a pure and separable bipartite state described by Def. 3.28. We know that the set of the quantum states D is a convex set and the set of the separable states S is also a convex set by definition. But this set (E) defined by $E \equiv D \setminus S$ is not a convex set[11]. It is obvious that $E \cap S = \emptyset$ and $E \cup S = D$. The Def. 3.28 and 3.29 can be extended for more than two parts. A n-part pure separable state can be written as a product of n pure states, $\rho = \rho^A \otimes \cdots \otimes \rho^n$ and a mixed separable state can be written as $\rho = \sum_i \lambda_i \rho_i^A \otimes \cdots \otimes \rho_i^n$.

Definition 3.30. *Every state* $\rho \in E$ *is an entangled state, i.e., if a state cannot be written as a separable state, then it is an entangled state*[12].

One could think that the entire space $\mathcal{H}_{\mathcal{AB}}$ can be generated by separable vectors $|u\rangle \otimes |v\rangle$ with $|u\rangle \in \mathcal{H}_{\mathcal{A}}$ and $|v\rangle \in \mathcal{H}_{\mathcal{B}}$. But there exists vectors that belong to $\mathcal{H}_{\mathcal{AB}}$ and that can not be written as a product state. In order to prove this fact, it is sufficient to exhibit one example. Then it is appropriate to introduce the *Bell states*, which are the simplest examples of entangled states [10]. These states are defined in Eq. 3.1 and 3.2 and form a basis for the 2-qubit state space:

$$|\Phi_\pm\rangle = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle), \tag{3.1}$$

$$|\Psi_\pm\rangle = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle). \tag{3.2}$$

It is obvious that $|\Phi_\pm\rangle$ and $|\Psi_\pm\rangle$ can not be written as product states of two particles [13], *i.e.;* $(|00\rangle \pm |11\rangle) \neq (a|0\rangle + b|1\rangle) \otimes (c|0\rangle + d|1\rangle)$ and $(|01\rangle \pm |10\rangle) \neq (a'|0\rangle + b'|1\rangle) \otimes (c'|0\rangle + d'|1\rangle)$, then, $|\Phi_\pm\rangle \neq |\phi_A\rangle \otimes |\phi_B\rangle$ and $|\Psi_\pm\rangle \neq |\psi_A\rangle \otimes |\psi_B\rangle$, but these Bell states are also states of $\mathcal{H}_{\mathcal{AB}}$. Then we can say that the four Bell states defined in 3.1 and 3.2 are entangled.

The natural question now is, "how many" states are entangled? In a more precise language, we intend to understand if the majority of the quantum states are separable or entangled. This discussion can be bounded by some volume calculations as done in reference [31], but here we just want to present

[10] Physicists like to simplify notations. All these notations can be found in physics textbooks $|1\rangle \otimes |0\rangle \equiv |1\rangle|0\rangle \equiv |1, 0\rangle \equiv |1\,0\rangle$.

[11] We can mix two elements $\rho_1, \rho_2 \notin S$, $\sigma = \lambda\rho_1 + (1-\lambda)\rho_2$, $0 \leq \lambda \leq 1$ and have $\sigma \in S$. An easy example is to write the trace one identity operator $\mathbb{I}_{4\times4}$, which is separable, by the following mixture $\mathbb{I}_{4\times4} = \frac{1}{4}\mathbb{I}_{2\times2} \otimes \mathbb{I}_{2\times2} = \frac{1}{2}\sum_B |B\rangle\langle B|$ where B represents all Bell states. The convexity of this mixture can be seen when we write in the $|i, j\rangle$ basis.

[12] Obviously this definition is not operational.

a brief explanation on this subject. For pure states, this answer is simple[13]: we have seen in Chapter 2 that we need 2 real numbers to describe one pure qubit in its Bloch sphere [20]. Then with the objective to describe two pure separable qubits, only 4 real numbers are necessary. But with the purpose to describe an element of $\mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$, (both complex spaces with dimension 2 each), we need 6 real numbers (2 complex numbers for each linear independent vector = 8), but we need just 6 numbers because 1 number is due to the global phase and 1 is due to the normalization [20, 31]. Since we need less dimensions to describe the pure and separable states, they look like a line in the $\mathcal{R}^3$ space and they form a thin set. However for mixed states, this is no longer true, the separable set for mixed states is also a dense set, then it is necessary to calculate the volume of these sets in order to quantify precisely their relative *size*, (see for example [31, 35–37]).

[13] Private communication with M. T. Quintino.

CHAPTER 4

Introduction to Classical Information Theory

> The most important questions of life are indeed, for the most part, really only problems of probability.
>
> Pierre-Simon, marquis de Laplace–Théorie Analytique des Probabilités, 1812.

4.1 Majorization and Partial Ordering

How do we compare two probability distribution vectors? What does it mean to say that one probability distribution vector is more disordered than another? To answer these kind of questions the concept of majorization was developed [38]. As in Classical Information theory, we also often encounter normalized vectors of non-negative numbers in Quantum Mechanics that can be interpreted as probability distribution vectors. If we possess a quantum state in hands, we can produce a probability distribution performing a set of measurements. We can compare as well two probability distribution vectors and two density matrices in a more elegant and efficient way and, for such purposes, and many others, the theory of majorization was developed.

Suppose we have a n-dimensional vector $\vec{x} = (x_1, x_2, \cdots, x_n)^t \in \mathcal{R}_+^n$, and also suppose that $\sum_{i=1}^n x_i = 1$. Then the set of all normalized vectors forms an $(n-1)$-dimensional simplex (Δ_{n-1}). We are interested here in transformations that preserve both positivity and the L_1-norm, ($\|\star\|_1$) of the vectors [31]. These transformations are given by the stochastic matrices as we will see later.

Given a probability vector $\vec{x} \in \mathcal{R}_+^n$, its decreasing rearrangement is denoted by[1] $\vec{x}^\bullet = (x_1^\bullet, x_2^\bullet, \cdots, x_n^\bullet)^t$, that is, $x_1^\bullet \geq x_2^\bullet \geq \cdots \geq x_n^\bullet$. A vector $\vec{x}$ is said to

[1] This strange, but beautiful "bullet" notation is due to [39].

be majorized by $\vec{y}$, in notation $\vec{x} \prec \vec{y}$, if:

$$\sum_{j=1}^{k} x_j^{\bullet} \leq \sum_{j=1}^{k} y_j^{\bullet}, \quad k = 1, 2, \cdots, (n-1),$$

$$\sum_{j=1}^{n} x_j^{\bullet} = \sum_{j=1}^{n} y_j^{\bullet}.$$

Definition 4.1 (Probability Simplex)**.** *A $(n-1)$-dimensional probability simplex can be defined as: $\vec{x} \in \mathcal{R}_+^n$ such that $\vec{x} \succeq \vec{0}$ and $\|\vec{x}\|_1 = 1$.*

4.1.1 Stochastic and the Bistochastic Maps

We know that the passage of time tends to make things more uniform and *pureless*. Thus, we need to understand the processes and their transformations that are responsible for this natural occurrence that brings the systems into the direction of the majorization arrow [31]. These maps are the stochastic and bistochastic maps and they appear in several physical problems. They are used in the theory of majorization (see [39]), and in characterization of completely positive maps acting in the space of density matrices [40]. In order to describe the discrete dynamics in the space of probabilities (or in the probability simplex) we need to define such maps (Definition 4.2 and 4.3).

Definition 4.2. *A stochastic matrix is a n-row rectangular matrix S whose matrix elements obey [31]:*

1. $S_{ij} \geq 0, \ \forall i, j.$
2. $\sum_{i=1}^{n} S_{ij} = 1.$

Definition 4.3. *A bistochastic matrix (also called doubly stochastic) is a n-dimensional stochastic matrix B obeying the additional condition:*

3. $\sum_{j=1}^{n} B_{ij} = 1.$

If the matrix $U = u_{ij}$, defined by its matrix elements is an unitary matrix, then the matrix B, defined as $B = (|u_{ij}|^2)$ is bistochastic. The condition 1. preserves the positivity of $\vec{x}$, *i.e.*, if $\vec{x}$ is a vector of $\mathcal{R}_+^n$ *i.e.*, if $\vec{x} \geq 0$, then $B\vec{x} \geq 0$. The condition 2. says that B preserves the norm $\|\star\|_1 \equiv tr(\star)$ when B acts on a positive vector $\vec{x}$. Let us define the $\vec{e}$ vector as $\vec{e} = (1, 1, \cdots, 1)^t$. The trace of a vector $\vec{x}$ is defined as[2] $tr(\vec{x}) \equiv \langle \vec{x}|\vec{e}\rangle$, then it is obvious to see that[3] if B is a bistochastic matrix, then $B\vec{e} = \vec{e}$, (this condition means that bistochastic matrices are *unital*). Then it is easy to understand the condition 3., or in other words, it is easy to show that $tr(B\vec{x}) = tr(\vec{x})$, since $tr(B\vec{x}) = \langle \vec{x}|B|\vec{e}\rangle$ and using the fact that the bistochastic matrices are unital, we have $\langle \vec{x}|B|\vec{e}\rangle = \langle \vec{x}|\vec{e}\rangle$, then $tr(B\vec{x}) = tr(\vec{x})$ [39].

[2] Roughly speaking, the trace function here is defined by the sum of the components of a vector, a kind of scalar product using the Dirac notation, *i.e.*, $tr(\star) = \langle \star|\vec{e}\rangle$.

[3] In Dirac notation $B\vec{e} = \vec{e}$ can be written as: $B|\vec{e}\rangle = |\vec{e}\rangle$.

4.1.2 Some Results in Majorization Theory

Theorem 4.4 (Birkhoff's Theorem). *The set of $n \times n$ bistochastic matrices is a convex polytope (Def. 3.17) whose pure points are the $n!$ permutation matrices.*

Proof: We already know that the bistochastic matrices are stochastic matrices which obey the additional condition $\sum_{j=1}^{n} B_{ij} = 1$. The permutation matrices Π_k are square and binary bistochastic matrices consisting of exactly one entry 1 in each row and each column and 0's elsewhere[4]. Let us suppose that any bistochastic matrix B can be written in a convex combination of these permutation matrices, *i.e.*, $B = \sum_k \lambda_k \Pi_k$, with $0 \leq \lambda_k \leq 1$ and $\sum_k \lambda_k = 1$, then:

$$B = \sum_k \lambda_k \Pi_k,$$

$$(B)_{ij} \overset{a}{=} \sum_k \lambda_k (\Pi_k)_{ij},$$

$$\sum_{i=1}^{n} B_{ij} = \sum_{j=1}^{n} B_{ij} \overset{b}{=} 1,$$

$$\sum_{i=1}^{n} (B)_{ij} = \sum_{i=1}^{n} \sum_k \lambda_k (\Pi_k)_{ij},$$

$$\sum_{i=1}^{n} (B)_{ij} \overset{c}{=} \sum_k \lambda_k \sum_{i=1}^{n} (\Pi_k)_{ij} \overset{d}{=} \sum_k \lambda_k = 1,$$

$$\sum_{j=1}^{n} (B)_{ij} \overset{e}{=} \sum_k \lambda_k \sum_{j=1}^{n} (\Pi_k)_{ij} \overset{f}{=} \sum_k \lambda_k = 1. \ \square$$

We start the proof supposing that any bistochastic matrix can be written in a convex combination of the permutation matrices. In a we exhibit the ij-th matrix element of B. The equality b are the bistochastic conditions (4.2 and 4.3). In c and e we just impose these conditions and in d and f we use the fact that the permutation matrices are also bistochastic matrices and the fact that $\sum_k \lambda_k = 1$.

Then it is easy to see that any bistochastic matrix B can be written as a convex combination of the permutation matrices. This convex combination obeys all the three conditions imposed in 4.2 and 4.3. Of course we should need to prove that the $n \times n$ permutation matrices are the pure points of this polytope and that the number of pure points needed in order to construct such convex mixture is $(n-1)^2$, by the Theorem 3.20, but this is completely out of the scope of this text.

Definition 4.5 (Muirhead's Condition). [26] *The Muirhead's condition states that if $\vec{x} \in P(\vec{y})$, which is a notation to express that there are nonnegative weights $p_i \geq 0$, with $\sum_i p_i = 1$, then $\vec{x} = \sum_{i \in P_n} p_i \Pi_i \vec{y}$, where the summation is performed over all possible permutations.*

Lemma 4.6. *The Muirhead's condition $\vec{x} \in P(\vec{y})$ implies that there exists a bistochastic matrix B such that $\vec{x} = B\vec{y}$.*

[4] It is a kind of binary sudoku, see http://xkcd.com/74/

Proof: We already show that $\sum_i p_i(\Pi_i)$ is a representation of a bistochastic matrix. Then we can say that the Muirhead's condition $\vec{x} \in P(\vec{y})$ implies that there exists a bistochastic matrix B such that $\vec{x} = B\vec{y}$.

Theorem 4.7 (HLP - Theorem[5]). [41] *For all $\vec{x}$ and $\vec{y}$ in $\mathcal{R}^n_+$, the following conditions are equivalent:*
i. $\vec{x} \prec \vec{y}$.
ii. $\vec{x} = B\vec{y}$. For some B bistochastic matrix.

Proof: Let us suppose that $\vec{x}^\bullet = (x_1^\bullet, x_2^\bullet, \cdots, x_n^\bullet)^t$, that is, $x_1^\bullet \geq x_2^\bullet \geq \cdots \geq x_n^\bullet$ [26]. Let us define also the sum of the first k elements of the t-th column of B, *i.e.*, $c_t = \sum_{j=1}^{k} B_{jt}$ [26]. Then:

$$\sum_{j=1}^{k} x_j = \sum_{t=1}^{n}(\sum_{j=1}^{k} B_{jt})y_t = \sum_{t=1}^{n} c_t y_t.$$

By the bistochasticity of B, we have $0 \leq c_t \leq 1$ and $\sum_{t=1}^{n} c_t = k$, because $\sum_{t=1}^{n} c_t = \sum_{t=1}^{n}\sum_{j=1}^{k} B_{jt} = \sum_{j=1}^{k}(\sum_{t=1}^{n} B_{jt}) = \sum_{j=1}^{k}(1) = k$. This strongly suggests that the k difference functions Δ_k defined below are non-positive functions:

$$\Delta_k \equiv \textstyle\sum_{j=1}^{k} x_j - \sum_{j=1}^{k} y_j = \sum_{t=1}^{n} c_t y_t - \sum_{j=1}^{k} y_j \leq 0.$$

A good and clever way to see that all functions Δ_k are non-positive (due to [26] or [42]) is to write the latter equation as:

$$\Delta_k \equiv \sum_{t=1}^{n} c_t y_t - \sum_{j=1}^{k} y_j = \sum_{j=1}^{n} c_j y_j - \sum_{j=1}^{n} y_j + y_k \times 0,$$

$$0 = (k - \sum_{j=1}^{n} c_j),$$

$$\Delta_k = \sum_{j=1}^{n} c_j y_j - \sum_{j=1}^{n} y_j + y_k(k - \sum_{j=1}^{n} c_j),$$

$$\Delta_k = \sum_{j=1}^{k}(y_k - y_j)(1 + c_j) + \sum_{j=k+1}^{n} c_j(y_j - y_k).$$

It is evident that $\Delta_k \leq 0$, since for all $1 \leq j \leq k$ we have $y_j \geq y_k$ while for all $k < j \leq n$ we have $y_j \leq y_k$. It is trivial that $\Delta_n = 0$, so the relations $\Delta_k \leq 0$ for $1 \leq k \leq n$ complete our check of the definition [26]. Therefore, if $\vec{x} = B\vec{y}$, then $\vec{x} \prec \vec{y}$ [41, 42]. □

Corollary 4.8. *A vector $\vec{x} \in \mathcal{R}^n_+$ is said majorized by a vector $\vec{y} \in \mathcal{R}^n_+$, i.e., $\vec{x} \prec \vec{y}$ iff there exists a set of n permutation matrices n-dimensional $\{\Pi_i\}_{i=1}^n$ and a probability distribution $\{p_i\}_{i=1}^n$ such that $\vec{x} = \sum_{i=1}^n p_i \Pi_i \vec{y}$* [38].

[5] Hardy, Littlewood and Polya - 1934

This implies that the vector $\vec{x}$ can be obtained from $\vec{y}$ by randomly permuting the components of $\vec{y}$, then averaging over the permutations [38].
Proof: This proof is simple because we have already showed that any bistochastic matrix can be written in a convex combination of permutation matrices (Theorem 4.4) and we have also showed in Lemma 4.6. Then the matrix $\tilde{B} \equiv \sum_{i=1}^{n} p_i \Pi_i$ is a bistochastic matrix. Hence by using Theorem 4.7, we obtain the expected result, *i.e.*, $\vec{x} = \tilde{B}\vec{y}$, which implies that $\vec{x} \prec \vec{y}$. □

Corollary 4.9. *The Muirhead's condition $\vec{x} \in P(\vec{Y})$ implies $\vec{x} \prec \vec{y}$.*

Proof: The proof comes from the fact of the equivalence of *i*. and *ii*. of the Theorem 4.7 and the other theorems of this section. The omitted parts can be found in [39].

Theorem 4.10. *Let be $\vec{x}_n = (\frac{1}{n}, \cdots, \frac{1}{n})^t$ the maximally mixed vector and $\vec{P} = (1, 0, \cdots, 0)^t$ the pure vector, then $\vec{x}_n \preceq \vec{x} \preceq \vec{P}$ for all $\vec{x} \in \mathcal{R}_+^n$* [31].

Proof: Intuitively, the vector $\vec{x}_n$ has the minimal differences between its elements, so all other vectors $\vec{x} \in \mathcal{R}^n$ majorize it. Let us observe the fact that $\vec{x} \prec \vec{y}$ *iff* there exists a bistochastic matrix B such as $\vec{x} = B\vec{y}$, and the uniform distribution vector or the maximally mixed vector defined above stays clearly invariant with respect to any bistochastic map. So we can say that the bistochastic map describes a contraction[6] of the probability simplex toward the uniform distribution [43]. Then there will always exist a bistochastic matrix $B = B_1 B_2 \cdots B_k$, (described here as a product of k bistochastic matrices), which makes this contraction possible: $\vec{x}_n = B\vec{y}$, for any $\vec{y} \in \mathcal{R}^n$.

It is obvious that the pure vector, $(\vec{P})$ majorizes every vector $\vec{x}$, or in other words, $\vec{x} \prec \vec{P}$ for all vector $\vec{x} \in \mathcal{R}^n$. The proof is simple since any of the n pure vector is constructed as a permutation of $\vec{P}$, then if we rearrange the vectors as usual, we will always have $\sum_{i=1}^{k} x_i^{\bullet} \leq \sum_{i=1}^{k} P_i^{\bullet} = 1$ and $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} P_i = 1$, thus $\sum_{i=1}^{k} x_i^{\bullet} \leq 1$. Hence, $\vec{x}_n \prec \vec{x} \prec \vec{P}$, for all $\vec{x} \in \mathcal{R}^n$. □

Definition 4.11 (Schur convex function)**.** *The function f which preserves the majorization order: if $\vec{x} \prec \vec{y}$, then $f(\vec{x}) \prec f(\vec{y})$ is called Schur convex* [44]*, a function f is Schur concave if $(-f)$ is Schur convex.*

[6] Of course the identity matrix is a bistochastic matrix, but $\mathbb{I}\vec{x} = \vec{x}$. If B is a bistochastic matrix, so any matrix $\mathbb{I} \oplus B$ is also a bistochastic matrix, however it preserves the subspace related to the identity matrix.

4.2 Shannon's Entropy $H(X)$

> You should call it entropy, for two reasons, Von Neumann told him. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, nobody knows what entropy really is, so in a debate you will always have the advantage.
>
> Suggestion of J. Von Neumann to C. Shannon on what to call information -version according to John Avery.

Suppose we have a probability distribution vector given by $\vec{p}$. We need to define a function $\mathcal{I} : (0, 1] \longmapsto \mathcal{R}_+$ that can measure the uncertainty of a random variable and quantify information of some events [45]. In 1948, Shannon modeled information as probabilistic events which can occur with certain probabilities [10, 46, 47]. A remarkable feature of Shannon approach is to ignore all the semantics and focuses on the statistical constraints that restrict the transmission of a message, regardless of its content [48]. Suppose we have a discrete random variable[7] X such as $p(x) = Pr(X = x)$. Shannon showed that any function $\mathcal{I}$ that can measure information must have these following properties:

1. The information of a probabilistic event must depend only on its probability, then $\mathcal{I} = \mathcal{I}(p)$.
2. The function $\mathcal{I} : (0, 1] \longmapsto \mathcal{R}_+$ must be a continuous function of p.
3. The function $\mathcal{I}$ must be additive, *i.e.*, if two events are independent[8] $p(x, y) = p(x)p(y)$, then $\mathcal{I}(p(x, y)) = \mathcal{I}(p(x)) + \mathcal{I}(p(y))$.

Shannon also showed in [47] that these three conditions above imply that $\mathcal{I}(p(x)) = -\log p(x)$ and this function is unique up to an affine transformation (see [46]). Shannon's entropy, thus, appears as the average missing information, that is, the average information required to specify the outcome X when a customer or receiver receives the distribution $p(X)$ [48]. It equivalently measures the amount of uncertainty represented by a probability distribution [49]. In our context here, it amounts to the minimal number of bits that should be transmitted to specify the variable X, but this concept needs to be wiredrawn with the data compression and coding theorems, see [10, 50]. For us, this definition is sufficient since we do not work with classical and quantum channel capacities.

Definition 4.12. *The expectation value of the random variable $g(X)$ if X occurs with probability $p(x)$ is defined by $E_{p(x)}(g(X)) \equiv \sum g(x)p(x)$.*

[7] The following notation will be used everywhere in this text: $P(x) = Pr(X = x)$, $P(y) = P(Y = y)$, $P(x, y) = Pr(X = x \,\&\, Y = y) = Pr(X = x \cap Y = y)$, $P(x|y) = Pr(X = x|Y = y)$.

[8] If two events are statistically independent, then $p(x, y) = p(x)p(y)$, *i.e.*, $Pr(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} Pr(A_i)$.

Definition 4.13. *The Shannon's entropy of X is the expected value of the random variable $g(X) = \mathcal{I}(p(x)) = \log(\frac{1}{p(x)})$, where X occurs with probability $p(x)$. Thus $H(X) \equiv E_{p(x)}(\log(\frac{1}{p(x)}))$, or in other words*[9]*, $H(X) = -\sum p(x)\log(p(x))$.*

Shannon's entropy $H(X)$ quantifies how much information is needed, *i.e.*, how many bits are required *on the average*, to encode n bits of information [50]. It is a measure of the uncertainty of the random variable that is; it is a measure of the amount of information needed, on the average, in order to describe a random variable [45]. Suppose we have a binary alphabet $X = \{x;\, p(x)\} = \{0,\, 1;\, p(0) = (1-p),\;\; p(1) = p\}$. If n is large enough, then by using the law of large numbers [45, 51], the typical strings will have approximately $n(1-p)$ letters 0 and np letters 1. The number of distinct strings is of the order of the Newton's binomial coefficient [50]:

$$\text{number} \approx \binom{n}{np},$$

By using Stirling's approximation [52], we have:

$$\log\binom{n}{np} = \log\frac{n!}{np!\,(n-np)!} \approx$$

$$\approx n\log n - n - [np\log np - np + n(1-p)\log n(1-p) - n(1-p)] =$$

$$= n[-p\log p - (1-p)\log(1-p)] \equiv nH(X).$$

Then the number of typical strings is about $2^{nH(X)}$ [50].

4.3 Some Properties of Shannon's Entropy

Lemma 4.14 (The entropy is non-negative)**.** $H(X) \geq 0$.

Proof. Since[10] $0 \leq p(x) \leq 1$, then $H(X) = \sum_{x\in X} p(x)\log(\frac{1}{p(x)}) \geq 0$. □

Theorem 4.15. *Let Π be a permutation matrix. Then the Shannon's entropy remains invariant under the map $\tilde{X} = \Pi X$, i.e., $H(\tilde{X}) = H(X)$ under permutations.*

Proof: The proof is quite simple. When we perform a permutation over the probability vector, the following function $-\sum_i p_i \log p_i$ must remain the same. This function $H(X)$ is not a feature of the random variable itself, but it is a intrinsic characteristic of the set of its probability values [45]. Of course if we re-label the indices we will find the same function for the two probability vectors. Therefore let us define a permutation matrix $\Pi_{\{(i+1),\,i\}}$ which, for the sake of simplicity, just changes the position of the i-th therm of X by the $(i+1)$-th therm of X, *i.e.*, it changes the order of p_i by the $p_{(i+1)}$. It is obvious that $H(\Pi_{\{(i+1),\,i\}}X) = \sum_{k=1}^{i-1} p_k \log p_k + p_{(i+1)}\log p_{(i+1)} + p_i \log p_i + \sum_{k=i+2}^{n} p_k \log p_k = H(X) = \sum_k p_k \log p_k$. □

[9] Remember that $\log(\frac{1}{p(x)}) = -\log(p(x))$.

[10] We use the following definition: $0\log 0 = 0$, for the sake of continuity, *i.e.*, $\lim_{x\to 0} x\log x = 0$.

Theorem 4.16 (Maxent Theorem - Principle of Maximum Entropy)**.** *For a normalized random variable with n possible values, the Shannon's entropy* $H(X) \leq \log n$*. The equality occurs iff when* $p(x) = \frac{1}{n}$, $\forall x$.

The theorem 4.16 follows easily from the concavity of the function entropy and from our intuition that the uncertainty is maximum when all events are equally probable [53]. But there are more rigorous ways of proving this theorem. $H(X)$ is a concave function, so it has a point of maximum. Then we could maximize it and get the proof. A more elegant way to prove this theorem is using the Jensen's inequality (Theorem 4.17).

Theorem 4.17 (Jensen's inequality)**.** *Let X be a random variable and f is a strictly convex function, then* $f(E(X)) \leq E(f(X))$*. If f is a strictly concave function, then* $f(E(X)) \geq E(f(X))$*. In more variables, let us suppose that* $X = \{x_i\}$ *is a random set, and let* $\sum_i a_i = 1$*, with* $a_i \geq 0$ $\forall i$*, and let f be a strictly* convex *function, then* $f(\sum_i a_i x_i) \leq \sum_i a_i f(x_i)$*, and if f is a strictly* concave *function, then* $f(\sum_i a_i x_i) \geq \sum_i a_i f(x_i)$.

Proof: We prove this inequality just for the binary case. The general proof is constructed by induction and the reader can find it in [45]. But it is easy to see that it works geometrically because generalizes the assertion that a secant line of a convex function lies above its graph. We need to answer first what is a convex and a concave function[11], see Def. 4.18 and 4.19 below:

Definition 4.18 (Convex function)**.** *A continuous function f is strictly convex over* (a, b) *if for every* $x_1, x_2 \in (a, b)$ *and* $\forall \lambda \in [0, 1]$, $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$*. In other words, f(the linear combination of* $\star$*)* $\leq$ *linear combination of* $f(\star)$.

Definition 4.19 (Concave function)**.** *A function f is said to be strictly concave if* $(-f)$ *is a convex function. Then f(the linear combination of* $\star$*)* $\geq$ *linear combination of* $f(\star)$.

Let us come back to the proof of the Jensen's inequality for the binary case[12]. Let $x \in \{x_1, x_2\}$ and $p(x_1) = p$, and $p(x_2) = 1 - p$. Let us suppose that f is a strictly convex function. We need to prove that if f is a convex function, then $f(E(X)) \leq E(f(X))$. Well, for any function f the expected value is defined as $E(f(X)) = \sum_x p(x) f(x) = pf(x_1) + (1-p) f(x_2)$. And the left side of the inequality is given by $f(E(X)) = f(\sum_x p(x)x) = f(px_1 + (1-p)x_2)$. By setting $p = \lambda \in [0, 1]$ and by using the definition of a convex function (Def. 4.18), we get the proof of the Jensen's inequality: **Proof - convex case:**

$$
\begin{gathered}
f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \text{ Def. 4.18,} \\
(f \text{ is a convex function and setting } p = \lambda), \\
f(E(X)) = f(px_1 + (1-p)x_2) \leq pf(x_1) + (1-p) f(x_2) = E(f(X)), \\
f(E(X)) \leq E(f(X)). \ \square
\end{gathered}
$$

[11] A good mnemonic rule is: "A function is *convex* if its graphic looks like a letter *V*, (of con*V*ex), and a function is *concave* if its graph looks like a *CAVE*, (of con*CAVE*)".

[12] The general proof is also performed by induction.

On the other hand, the proof of Jensen's inequality for concave functions is identical. Then if f is a concave function we will have: **Proof - concave case:**

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2), \text{ Def. 4.19,}$$
$$(f \text{ is a concave function and setting } p = \lambda),$$
$$f(E(X)) = f(px_1 + (1-p)x_2) \geq pf(x_1) + (1-p)f(x_2) = E(f(X)),$$
$$f(E(X)) \geq E(f(X)). \square$$

We can prove the Theorem 4.16 in a more elegant fashion by using the Jensen's inequality (Theorem 4.17) in the concave version because $f = \log(\star)$, in this case, it is a concave function:

$$\text{Let } H(X) = -\sum_{x=1}^{n} p(x)\log(p(x)),$$
$$E(f(X)) \leq f(E(X))$$
$$H(X) = \sum_{x=1}^{n} p(x)\log\left(\frac{1}{p(x)}\right)$$
$$H(X) \leq \log\left(\sum_{x=1}^{n} p(x)\frac{1}{p(x)}\right)$$
$$H(X) \leq \log n. \square$$

4.3.1 The Bayes' Rule

Let us understand the Bayes' Rule with the help of the Venn's diagrams. In a universe Ω with all possible outcomes, we are interested in a subset A of Ω. We know that the probability of obtaining the outcome A is: $P(A) = \frac{|A|}{|\Omega|}$, where the $|\star|$ denotes, for example, the area occupied in the chosen set. Then, obviously, $P(\Omega) = \frac{|\Omega|}{|\Omega|} = 1$, because it is equivalent to the total area.

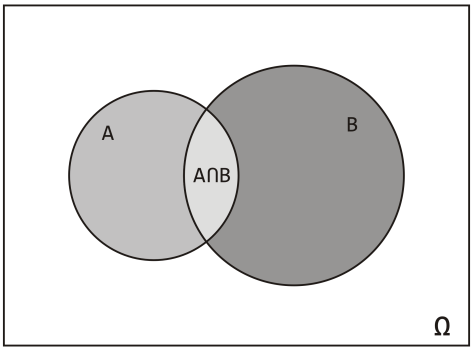

Figure 4.1: A two-set Venn's diagram involving A, B, $A \cap B$ and the universe Ω.

Let us consider another event in the space Ω, the probability $P(B) = \frac{|B|}{|\Omega|}$. We have treated these two events independently, but of course these two events can be dependent ($A \cap B = \emptyset$), see Fig 4.1. The probability of both

events occurrence, *i.e.*, $P(A \cap B)$, can be calculated by using the same idea: $P(A,\ B) \equiv P(A \cap B) \equiv P(A\&B) = \frac{|A,B|}{|\Omega|}$.

Now, the question is: given that we are inside the area B, what is the probability that we are in the region $A \cap B$? We need to calculate the area of A if we make $B = \Omega_B$ our new universe (see Fig. 4.2). This is usually called a conditional probability $P(A|B)$ and we read: the probability of A given B [14].

Then using the new universe definition we have[13] $P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A,B|}{|B|}$. But we can divide both the numerator and the denominator by the total area $|\Omega|$, and substitute the previous relations defined above, $P(A|B) = \frac{\frac{|A,B|}{|\Omega|}}{\frac{|B|}{\Omega}}$. Then $P(A|B) = \frac{P(A,B)}{P(B)}$.

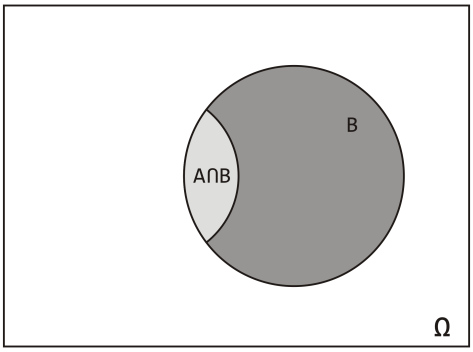

Figure 4.2: A two-set Venn's diagram involving B, $B - A$ and the universe Ω.

Of course this reasoning would be the same if we consider our new universe, the set A. Then it is obvious that $P(B|A) = \frac{P(A,B)}{P(A)}$. If we put these two last equations together, we will have the Bayes' rule:

Definition 4.20 (Bayes' Rule)**.** $P(A|B)P(B) = P(A,\ B) = P(B|A)P(A)$.

$$P(A,\ B) = P(A)P(B|A),$$
$$P(A,\ B) = P(B)P(A|B).$$

4.3.2 Markov Chains and Bistochastic Maps

The objective followed in this subsection[14] is to discuss briefly some properties of the Markov chains to use further in order to simplify some proofs of a few theorems and also to introduce the entropy increasing theorem and some of its relationships with the Second Law of Thermodynamics.

We also intend to understand the stochastic maps that act on the probability vectors in the classical probability space. Later this concept will be useful when

[13] From now on we are going to use the notation $A \cap B \equiv A,\ B$.
[14] This spirit is also followed in [10].

we try to understand the quantum states as quantum probability distribution entities of the quantum probability space. If we want to model noise in quantum systems, we should study first the classical noise [10].

The simplest example of a stochastic process is one where each random variable is conditionally independent of all other preceding variables and depends only those which precede them, *i.e.*, given the present time, the future is independent of the past [45]. Such a process is said to be Markov. Physically, the assumption of Markovianity corresponds to assuming that the environment causing the noise in a gate X acts independently of the environment causing the noise in gate Y, (if we model a physical process as stochastic process consisting on two gates: $X \to Y$ [10]).

We will assume that all the Markov chains studied here are time invariant since the following conditional probability $p(x_n|x_{n-1})$ does not depend on n [45].

Definition 4.21 (Markov chain)**.** *A discrete stochastic process is a Markov Chain for* $n = 1, 2, \cdots$, *if* $Pr(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}|\cdots|X_1 = x_1) = Pr(X_{n+1} = x_{n+1}|X_n = x_n)$. *In this case, the joint probability is defined as* $p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1})$. *Two good notations for a finite Markov Chain are:*

- $X_1 \to X_2 \to \cdots \to X_{n-1} \to X_n$.
- $\{X_i\}$.

If $\{X_i\}$ is a Markov chain, thus, X_n is defined as its state at time n. A time-invariant Markov chain is defined and characterized by its initial state and a probability transition matrix $S = (S)_{ij}$, $\forall i, j \in \{1, 2, \cdots, m\}$, where $S_{ij} = Pr(X_{n+1} = j|X_n = i)$ [45]. Suppose that we want to know the probability vector on time $n+1$, then $p(x_{n+1}) = \sum_{x_n} S_{\{x_n, x_{n+1}\}} p(x_n)$. Thus the output probabilities are related to the input probabilities by a linear process mapped in the probability transition matrix. This feature of linearity is echoed in the description of quantum noise, with density matrices replacing probability distributions [10]. If the state at time $n+1$ is the same of the state at time n, then it is called a stationary distribution. Suppose that $\mu = S\mu$, then μ is a stationary distribution.

Theorem 4.22. *The transition matrix of a Markov Chain is stochastic.*

Proof: Since the probability of transitioning from state i to some state must be 1, we have that this matrix is a stochastic matrix, $S_{ij} = Pr(X_{n+1} = j|X_n = i)$, then $\sum_j S_{ij} = \sum_j Pr(X_{n+1} = j|X_n = i) = 1$. □

Theorem 4.23. *Let* $p(x_0), \cdots, p(x_n)$ *be a sequence where* $p(x_{n+1}) = Bp(x_n)$, *where* B *is a matrix. If* $p(x_{n\to\infty}) \to \mu = (\frac{1}{n}, \cdots, \frac{1}{n})$, *then* B *is a bistochastic matrix.*

Proof: Since $p(x_{n+1}) = Bp(x_n)$ the following holds $\forall j = 1, \cdots, n$:

$$p(x_{n+1})_{(j)} = \sum_{i=1}^{n} B_{ij} p(x_n)_{(i)},$$

$$p(x_{n+1})_{(j)} = \frac{1}{n} \sum_{i=1}^{n} B_{ij} = \frac{1}{n}.$$

The last reduction stems from the fact that each column sums to 1. As the value at each j-th index becomes $\frac{1}{n}$, this distribution is obviously a stationary distribution. The Fundamental Limit Theorem of Markov Chains, (see [54]), states that there is a unique stationary distribution for any probability transition matrix, therefore, the uniform distribution $(\frac{1}{n}, \cdots, \frac{1}{n})$, must be the unique stationary distribution for all bistochastic (or doubly stochastic matrices[15]), and any initial $p(x_0)$ must converge to $\mu = (\frac{1}{n}, \cdots, \frac{1}{n})$ [55].

We need to define more quantities before discussing some important theorems on stochastic maps and its relationships with the Markovian processes. The important issues are the relations of the bistochastic maps, the decreasing of the classical relative entropy and the entropy increasing theorem. But we need to study first what each concept means with detail.

4.3.3 Conditional Entropy $H(X|Y)$

If the considered probability vector is given by $p(x, y)$, then the conditional entropy is $H(X|Y) \equiv -E_{p(x,y)}(\log(p(x|y)))$.

Definition 4.24 (Conditional Entropy). $H(X|Y) = -\sum_{y\in Y}\sum_{x\in X} p(x, y)\log p(x|y)$.

Definition 4.25 (Conditional Entropy). $H(X|Y) = -\sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)\log p(x|y)$.

The Def. 4.25 shows the positivity of $H(X|Y)$, because it is a convex combination of another entropy, therefore $H(X|Y) \geq 0$, see [45, 53].

Theorem 4.26 (Chain Rule). $H(X, Y) = H(X) + H(Y|X)$.

Proof: Starting with $p(x, y) = p(x)p(y|x)$, taking the logarithm of both sides and later taking the expectation of both sides and remembering that $(\sum_{x\in X}\sum_{y\in Y} p(x, y) = \sum_{x\in X} p(x))$, we get the proof.

$$\begin{aligned}
p(x, y) &= p(x)p(y|x),\\
\log p(x, y) &= \log p(x) + \log p(y|x),\\
-E_{p(x,y)}(\log p(x, y)) &= -E_{p(x,y)}(\log p(x)) + (-E_{p(x,y)}(\log p(y|x)),\\
-\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x, y) &= -\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(y|x),\\
H(X, Y) &= H(X) + H(Y|X). \ \square
\end{aligned}$$

Remark: It is easy to see that in general $H(X|Y) \neq H(Y|X)$, but we will see later that $H(X) - H(X|Y) = H(Y) - H(Y|X)$. For a numerical example for this theorem, see [45].

[15] Again we need to exclude some obvious cases, as the identity matrix and $\mathbb{I} \oplus B$ and so on.

4.3.4 The Joint Entropy $H(X, Y)$

The definition of entropy of more than one variable is straightforward. This fact occurs because this new vector (X, Y) can be considered as a single random variable [45]. Let be a pair of discrete random variables (X, Y) with a joint probability distribution given by $p(x, y)$. Then the Joint entropy is defined by $H(X, Y) \equiv -E_{p(x,y)}(\log p(x, y))$, or in other words:

Definition 4.27 (Joint Entropy). $H(X, Y) = -\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x, y)$.

The *information* content of a joint distribution cannot be larger than the *sum* of the information contents of the individual distributions. We can state this phrase precisely with the help of the following theorem:

Theorem 4.28 (The Shannon's Entropy is sub-additive). $H(X, Y) \leq H(X) + H(Y)$. *The equality holds iff X and Y are statistical independent variables.*

Proof: We divide the proof in two cases. Let us prove first the equality (Case 1). When two variables are statistically independent, then $p(x, y) = p(x)p(y)$. If we take the logarithm and the expectation of both sides, we obtain the proof of the equality. In the Case 2, we suppose that X is not independent of Y.

Case 1: X and Y are independent random variables:

$$\begin{gathered}
p(x, y) = p(x)p(y), \\
\log p(x, y) = \log p(x) + \log p(y), \\
-E_{p(x,y)}(\log p(x, y)) = -E_{p(x,y)}(\log p(x)) + (-E_{p(x,y)}(\log p(y)), \\
-\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x, y) = -\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(y), \\
H(X, Y) = H(X) + H(Y). \square
\end{gathered}$$

Case 2: X is not independent of Y. The Shannon's Entropy is sub-additive: $H(X, Y) \leq H(X) + H(Y)$, because X and Y *are not* independent random variables. First observe that $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$, then: **Proof:**

$$\begin{aligned}
H(X) + H(Y) &= -\left(\sum_{x\in X} p(x)\log p(x) + \sum_{y\in Y} p(y)\log p(y)\right), \\
H(X) + H(Y) &= -\left(\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x) + \sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(y)\right), \\
H(X) + H(Y) &= -\left(\sum_{x\in X}\sum_{y\in Y} p(x, y)\log(p(x)p(y)\right),
\end{aligned}$$

On the other hand, $H(X, Y) = -\sum_{x\in X}\sum_{y\in Y} p(x, y)\log p(x, y)$. Combining these two last equations (and using the Jensen's inequality, Theorem 4.17 in

the inequality a, we obtain:

$$H(X, Y) - H(X) - H(Y) = \sum_x \sum_y p(x, y)\left(\log\left(\frac{1}{p(x, y)}\right) + \log(p(x)p(y))\right),$$

$$= \sum_x \sum_y p(x, y) \log\left(\frac{p(x)p(y)}{p(x, y)}\right) \overset{a}{\leq} \log\left(\sum_x \sum_y \frac{p(x, y)p(x)p(y)}{p(x, y)}\right),$$

$$H(X, Y) - H(X) - H(Y) \leq \log\left(\sum_x p(x) \sum_y p(y)\right) = 0,$$

$$H(X, Y) - H(X) - H(Y) \leq 0,$$

$$H(X, Y) \leq H(X) + H(Y). \square$$

Corollary 4.29 (Conditioning reduces entropy)**.** *$H(X|Y) \leq H(X)$ and the equality holds iff X and Y are independent.*

Proof: We have already shown that $H(X, Y) \leq H(X) + H(Y)$, (Theorem 4.28), with equality *iff* the variables are independent. By using Theorem 4.26, we have $H(X, Y) = H(Y) + H(X|Y) \leq H(X) + H(Y)$. Thus the corollary follows immediately, that is $H(X|Y) \leq H(X)$, and the equality holds *iff* the two variables are independent. $\square$ Another way to understand this theorem is to remember this definition: $H(X|Y = y) = \sum_{y \in Y} p(y)H(X|Y = y) \leq H(X)$. Of course learning the variable Y cannot reduce our knowledge of the variable X, in fact the opposite occurs: learning another variable can only increase our prior knowledge of some other variable, then conditioning can only reduce the entropy.

This Corollary could be proved in a more simple manner with the definition of mutual information in hands, in terms of relative entropy, (Subsection 4.3.5), and this proof would follow these steps: First, we prove that the mutual information is non-negative, by using the non-negativity of the relative entropy then, by using one of the definitions of mutual information that we obtain the proof.

More Than One Variable

Theorem 4.30. *For more than one variable, the Joint Entropy of the set $\{X_1, X_2, \cdots, X_n\}$ is $H(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} H(x_i|x_{i-1}, \cdots, x_n)$.*

Proof: Observe that $p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \cdots, x_1)$, taking the logarithm of both sides and then taking the expectation $-E_{p(x_1, x_2, \cdots, x_n)}$ of

both sides of this equation, we have:

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \cdots, x_1),$$
$$H(x_1, x_2, \cdots, x_n) = H(x_1) + H(x_2|x_1) + H(x_3|x_2, x_1) + \cdots + H(x_n|x_{n-1}, \cdots, x_1),$$
$$H(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} H(x_i|x_{i-1}, \cdots, x_n). \square$$

4.3.5 Shannon's Mutual Information function $I(X, Y)$

The Shannon's mutual information function quantifies the amount of information contained in a random variable about another random variable [53].

Definition 4.31. $I(X, Y) = \sum_{x\in X}\sum_{y\in Y} p(x, y)\log(\frac{p(x, y)}{p(x)p(y)})$.

Obviously the mutual information is a symmetric function of the two variables. We find out as much about the variable X by learning Y as well as learning Y through X [50]. When the two distributions are independent, $I(X, Y) = 0$. In the case where X and Y are the source and the output of a communication channel, the quantity $I(X, Y)$ measures the amount of information going through the channel. This amount cannot exceed the information of the source or that of the output [53]. Therefore, $I(X, Y) \leq H(X)$ and $I(X, Y) \leq H(Y)$. In terms of the conditional entropy and the classical relative entropy, we can define the mutual information as:

$$I(X, Y) \equiv H(X) - H(X|Y),$$
$$I(X, Y) \equiv H(Y) - H(Y|X),$$
$$I(X, Y) \equiv H(X) + H(Y) - H(X, Y).$$

The mutual information cannot be negative. This property means physically that learning X can never reduce our knowledge of Y and *vice versa* [50]. It is easy to see because $H(X|Y) \leq H(X)$, so $I(X, Y) \equiv H(X) - H(X|Y) \geq 0$. This is also true for $I(X, Y) \equiv H(Y) - H(Y|X) \geq 0$ because $H(Y|X) \leq H(Y)$.

4.3.6 Shannon's Entropy and Venn's Diagrams

In the following figure, (Fig. 4.3, [10]), we display a pictorial representation of the Shannon entropies $H(X)$ and $H(Y)$, the Joint Entropies $H(X, Y)$ and $H(Y, X)$. Also we have the Conditional entropies $H(X|Y)$ and $H(Y|X)$ and the Mutual Information $I(X, Y)$ [10, 50]:

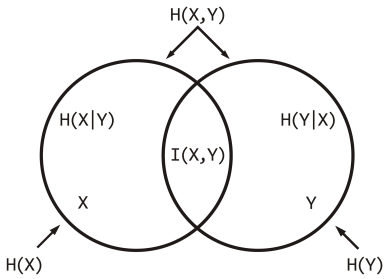

Figure 4.3: Shannon's entropy and Venn's diagrams.

4.4 The Classical Relative Entropy $D(P||Q)$

In information theory, we often need to measure and quantify differences between two probability distributions P and Q. The classical relative entropy function was introduced to quantify the *divergence* between a so called *true* distribution P and another distribution Q that, in general, is an approximation, an estimate or a model of P. Then it is a measure of the inefficiency of assuming the distribution Q instead using P [45]. However, we see that despite the classical relative entropy provide us this intuitive concept of relative *distance* between two probability vectors, it is not a metric function[16]. This function $D(P||Q)$ is often called *the Kullback–Leibler divergence function, or distance function or classical relative entropy*.

Definition 4.32 (Classical Relative Entropy). $D(P||Q) = E_{p(x)}(\log \frac{p(x)}{q(x)})$,

$$D(P||Q) = \sum_{x\in X} p(x) \log \frac{p(x)}{q(x)}.$$

Theorem 4.33. *Let P and Q be two n-dimensional probability distributions, then*[17] $D(P||Q) \geq 0$.

Proof: We are going to use the following inequality for the logarithm function: $\ln x \leq x - 1$, and the equality holds *iff* $x = 1$ [56]. Let us call $x = \frac{q(x)}{p(x)}$. Then:

$$-\sum_{x\in X} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x\in X} p(x)\left(\frac{q(x)}{p(x)} - 1\right),$$

$$-\sum_{x\in X} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x\in X} q(x) + 1,$$

[16] It is obvious that a huge problem happens when we are trying to compare two probability distributions and one is defined by $Q = \{1, 0, \cdots, 0\}$. Another consideration is that $D(P||Q)$ is often not equal to $D(Q||P)$, then the relative entropy is usually not symmetric. These are some of the reasons why the relative entropy can not be considered as a metric function.

[17] The Gibb's inequality is a special case of this theorem. Since $D(P||Q) \geq 0$, then $\sum_{x\in X} p(x) \log p(x) \geq \sum_{x\in X} p(x) \log q(x)$.

$$\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \geq -1+1,$$
$$D(P||Q) \geq 0. \square$$

It is easy to see that the equality in the theorem 4.33 holds *iff* $x = 1$, *i.e.*, if $x = \frac{q(x)}{p(x)} = 1$, which implies that $p(x) = q(x)$ for all x, or in other words: $P = Q$ [10]. Then $D(P||Q)$ is always non-negative[18].

We could define the mutual information as $I(X, Y) \equiv D(p(x, y)||p(x)p(y))$. We have already shown that $D(p(x, y)||p(x)p(y)) \geq 0$, and the equality holds *iff* X and Y are statistical independent, *i.e.*, $p(x, y) = p(x)p(y)$, thus the mutual information is also a non-negative quantity.

The following Theorem will be useful later in Section 4.5, when we try to understand the Second Law of Thermodynamics for Markovian processes.

Theorem 4.34 (The Chain Rule for the Relative Entropy)**.** $D(P(x, y)||Q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

Proof:

$$D(P(x, y)||Q(x, y)) = \sum_x\sum_y p(x, y)\log\left(\frac{p(x, y)}{q(x, y)}\right),$$
$$D(P(x, y)||Q(x, y)) = \sum_x\sum_y p(x, y)\log\left(\frac{p(x)p(y|x)}{q(x)q(y|x)}\right),$$
$$D(P(x, y)||Q(x, y)) = \sum_x\sum_y\left[p(x, y)\log\left(\frac{p(x)}{q(x)}\right) + p(x, y)\log\left(\frac{p(y|x)}{q(y|x)}\right)\right],$$
$$D(P(x, y)||Q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \square$$

4.4.1 The Classical Relative Entropy Means Something

Theorem 4.35 (Sanov's Theorem - Binary Case)**.** *Let an experiment be described by the following probability distribution* $Q = \{q, (1-q)\}$*, and let it be repeated* $\mathcal{N}$ *times. Let* $E \subset S$ *be a set of probability distributions, such is the closure of its interior, and let* S *be a probability simplex. Then, for large* $\mathcal{N}$*, the probability* $\mathcal{P}$ *of the following frequency distribution* $P = \{\frac{m}{\mathcal{N}}, (1-\frac{m}{\mathcal{N}})\}$ *belonging to* E*, is obtained by* $P(E) \approx e^{-\mathcal{N}D(P_*||Q)}$*, where* P_* *is the distribution in* E*, which has the smallest value of* $D(P||Q)$*, i.e.,* $P_* = argmin_{\{P\in E\}} D(P||Q)$ [31].

The proof needs some strong results of the large deviation theory[19] and it is beyond of the scope of this introductory text. But the reader can find these results, including the proof of the law of large numbers, in the Chapter 11 of [45]. But following the simple exercise discussed in [31], we can understand the essence of the Theorem 4.35 and try to figure its proof. In fact, we consider just

[18] A more elegant (and easier) way to proof the theorem 4.33 is to apply the Jensen's inequality for concave functions, with $x = \frac{q(x)}{p(x)}$, and with $f = \log(\star)$, (Theorem. 4.17). Then, $E_{p(x)}(f(\frac{q(x)}{p(x)})) = -D(P||Q) \leq f(E_{p(x)}(\frac{q(x)}{p(x)})) = \log(\sum(p(x)\frac{q(x)}{p(x)}))$, then $-D(P||Q) \leq \log 1 = 0 \Rightarrow D(P||Q) \geq 0. \square$

[19] The theory of large deviations concerns the asymptotic behaviour of sequences of probability distributions, see [45].

the binary case of this theorem simply for this reason. The set of frequencies $P = \{\frac{m}{\mathcal{N}}, (1 - \frac{m}{\mathcal{N}})\}$ is a matter of counting how many strings of outcomes exist with each given frequency [31]. We know how to calculate the probability $\mathcal{P}(\frac{m}{\mathcal{N}})$ by using the Binomial theorem and the Stirling's approximation for large $\mathcal{N}$:

$$\mathcal{P}(\tfrac{m}{\mathcal{N}}) = (\tbinom{\mathcal{N}}{m} q^m (1-q)^{\mathcal{N}-m},$$
$$\ln(\mathcal{P}(\tfrac{m}{\mathcal{N}})) \approx -\mathcal{N}\{\tfrac{m}{\mathcal{N}}[\ln(\tfrac{m}{\mathcal{N}}) - \ln(q)] + (1 - \tfrac{m}{\mathcal{N}})[\ln(1 - \tfrac{m}{\mathcal{N}}) - \ln(1-q)]\},$$
$$\mathcal{P} \approx e^{-\mathcal{N} D(P||Q)},$$
$$\text{Sanov's Theorem: } \mathcal{P}(E) \approx e^{-\mathcal{N} D(P^*||Q)}.$$

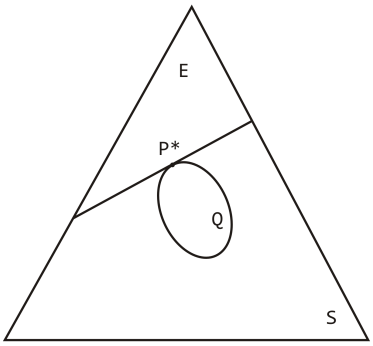

Figure 4.4: An example of Sanov's Theorem: the probability simplex with the probability distributions P^* and Q.

Of course $D(P_*||Q)$ can be a positive number in the convex set E, (which is, in this case, a triangle, see Fig. 4.4, [45]). If $D(P_*||Q) > 0$, then the only way to increase the probability is increasing the number of attempts $\mathcal{N}$. In this figure, we can see a pictorial representation of the optimal value P^* attained in the convex set E. The distance between P^* and Q, in this case, is non-zero. But the presented case in the binary exemplification of the Theorem 4.35 is too simple. As soon as the number of attempts increases, the $P_* \to Q$, by using the weak formulation of the law of large numbers, that is $\left|\frac{m(\mathcal{N})}{\mathcal{N}} - q\right| \to 0$, when $\mathcal{N} \to \infty$, then when $\mathcal{N}$ is large enough[20] the probability is near to 1.

Theorem 4.36 (The Minimum Relative Entropy Theorem)**.** *Let P be a normalized distribution and let $Q = \{\frac{1}{n}\}$ be a normalized homogeneous distribution. Then the principle of maxent with the normalization as the unique constraint (4.16), is equivalent to the minimization of the classical relative entropy of P relative to the homogeneous distribution, i.e., $max_X\{H(X),\ \sum p(x) = 1\}$ is identical to $min_X\{D(P||Q),\ \sum p(x) = \sum q(x) = \sum \frac{1}{n} = 1\}$.*

[20] The magic here is the exponential convergence of the sequences granted by the theory of large deviations.

Proof: The relative entropy is a convex function. Then it has a point of minimum, then $\delta[D(P||Q)] = 0$ and $\delta^2[D(P||Q)] > 0$.

$$\delta[D(P||Q)] = 0,$$
$$\delta[D(P||Q)] = \delta[\sum_{x\in X} p(x)\log(p(x))] + \delta[\ln n \sum_{x\in X} p(x)],$$
$$\delta[D(P||Q)] = -\delta[H(X)],$$
$$\delta^2[D(P||Q)] > 0 \Rightarrow \delta^2[H(X)] < 0. \ \square$$

Then it is true that:

$$max_{\{p(x)\in X\}}\{H(X),\ \sum p(x) = 1\} \text{ is equivalent to}$$
$$min_{\{p(x)\in X\}}\{D(P||Q),\ \sum p(x) = \sum \tfrac{1}{n} = 1\}.$$

Remark: The uniqueness of the solution of both theorems (maximizing the entropy 4.16) and (minimizing the relative entropy between the given distribution and the one which has the maximum entropy 4.36) is guaranteed because they are equivalent convex optimization problems over a closed compact set and thus they yield a unique solution [57].

When there exists more constraints than the normalization one, we can use the *maxent* principles by adding some proper constraints. But the Minimum Relative Entropy Theorem can not easily be used, because this principle assumes that there is one *true* probability distribution or model.

When there exists constraints, the only guarantee is that the probabilities are treated equally, without any bias, thus this claims a maximization of the entropy in a constrained set. But, of course, it does not imply that all probabilities will be equal to $\frac{1}{n}$. In fact, when we have constraints this solution is not, in general, compatible with the constraints[21].

The strength of the principle of minimum relative-entropy is precariously dependent on the knowledge of the *true* probability distribution, and it provides any mathematical meaning to communicate the amount of faith we have in it [57]. In Section 4.6, this issue is addressed properly.

4.4.2 A Geometrical Significance of The Classical Relative Entropy

We already know that the classical relative entropy $D(P||Q)$ cannot be considered as a metric function. It has some properties that a good candidate of metric function must have, but it is not symmetric. We have already proved that the classical relative entropy is a non-negative quantity 4.33. In fact $D(P||Q) \geq D_2^2(P,\ Q)$, where $D_2^2(P,\ Q) \equiv \frac{1}{2}\sum_{i=1}^{n}(p_i - q_i)^2$ [31].

Theorem 4.37. $D(P||Q) \geq D_2^2(P,\ Q)$.

[21] Special thanks to all folks from InfoQuant ($|IQ\rangle$) for all those *lovely* discussions about this fact.

Proof: (Due to [31]). Every smooth function obeys: $f(x) = f(y) + (x - y)f'(y) + \frac{1}{2}(x-y)^2 f''(\xi)$, with $\xi \in (x, y)$ [31]. Setting[22] $f(x) = -x \ln x$, with $f''(\xi) \leq -1$, when $0 \leq \xi \leq 1$, and with a rearrangement of terms we prove the theorem:

$$\begin{gathered}
f(x) = f(y) + (x-y)f'(y) + \tfrac{1}{2}(x-y)^2 f''(\xi), \\
-x\log x = -y\log y - (x-y)(\log y + 1) + \tfrac{1}{2}(x-y)^2 f''(\xi), \\
-x\log x = -x\log y - (x-y) + D^2(x,\, y)f''(\xi), \\
-x\log x + x\log y = -(x-y) + D^2(x,\, y)f''(\xi), \\
-D(x||y) = -(x-y) + D^2(x,\, y)f''(\xi), \\
D(x||y) \overset{a}{=} (y-x) - D^2(x,\, y)f''(\xi), \\
D(x||y) \overset{b}{\geq} c + D^2(x,\, y), \\
D(x||y) \geq D_2^2(x,\, y).
\end{gathered}$$

In a, we put $max_{\{0\leq\xi\leq 1\}} f''(\xi) = -1$ and in b, we defined $(y - x) = c > 0$.

Theorem 4.38 (Pythagorean Theorem)**.** [45] *Let $E \subset S$ be a convex set, where S is the probability simplex. Let $P \in E$ be a distribution and let $Q \notin E$ be another distribution. If $P_* \in E$ is the distribution that achieves the minimum distance to Q; that is, $D(P_*||Q) = min_{\{P\in E\}} D(P||Q)$. Then $D(P||Q) \geq D(P||P_*) + D(P_*||Q)$, for all $P \in E$.*

Proof: (Also due to [45]). Consider a distribution $P \in E$. Let $P_\lambda = \lambda P + (1-\lambda)P_*$. Of course $P_\lambda \to P_*$ when $\lambda \to 0$. As $E \subset S$ is a convex set and P_λ is constructed as a convex mixture, all $P_\lambda \in E\ \forall\lambda$. Since $D(P_*||Q)$ is the minimum of $D(P_\lambda||Q)$ along the path $P_\lambda \to P_*$, the derivative of $\frac{d}{d\lambda}(D_\lambda(P_\lambda||Q))|_{\lambda=0} \geq 0$. Then:

$$\frac{d}{d\lambda}(D_\lambda(P_\lambda||Q))|_{\lambda=0} = \sum_x \left[(P(x) - P_*(x))\ln\left(\frac{P_*(x)}{Q(x)}\right) + (P(x) - P_*(x))\right] \geq 0.$$

Using the fact that $\sum_x P(x) = \sum_x P_*(x) = 1$,

$$\begin{gathered}
\sum_x \left[(P(x) - P_*(x))\ln\left(\frac{P_*(x)}{Q(x)}\right)\right] \geq 0, \\
\sum_x \left[P(x)\ln\left(\frac{P_*(x)}{Q(x)}\right) - P_*(x)\ln\left(\frac{P_*(x)}{Q(x)}\right)\right] \geq 0, \\
\sum_x \left[P(x)\ln\left(\frac{P(x)P_*(x)}{Q(x)P(x)}\right) - P_*(x)\ln\left(\frac{P_*(x)}{Q(x)}\right)\right] \geq 0, \\
\sum_x \left[P(x)\ln\left(\frac{P(x)}{Q(x)}\right) - P(x)\ln\left(\frac{P(x)}{P_*(x)}\right) - P_*(x)\ln\left(\frac{P_*(x)}{Q(x)}\right)\right] \geq 0, \\
D(P||Q) - D(P||P_*) - D(P_*||Q) \geq 0, \\
D(P||Q) \geq D(P||P_*) + D(P_*||Q). \ \square
\end{gathered}$$

[22] From now on, we will use ln instead of log in the definition of $D(P||Q)$, and we will completely omit the factor of conversion.

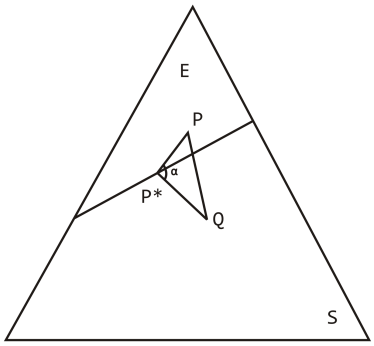

Figure 4.5: The Pythagorean theorem.

The Pythagorean theorem (4.38) shows an interesting property of $D(\star||\star)$. It shows that the classical relative entropy behaves like the square of the Euclidean distance.

Note that it is easier to understand this theorem by consulting the Figure 4.5, [45]. We have three triangles here: The convex subset E of the probability simplex S, which is, in this case, a triangle, the simplex S itself is a triangle and we can make a drawing of a triangle with vertices P, P_*, Q. Fixed a point $P \in E$, then we have another distribution $P^* \in E$, in such a way it achieves the minimum distance to the distribution Q. These distances, obviously, form a triangle.

Using a certain type of *triangle inequality* and the fact that $D(\star||\star) \sim d^2$, where d^2 is the Euclidean distance between the two distributions (distance between two n-dimensional vectors in $\mathcal{R}_+^n$), and by using Theorem 4.38, we can write $d^2(P, Q) \geq d^2(P, P_*) + d^2(P_*, Q)$. Then the angle $\alpha \equiv \angle(PP_*, P_*Q)$ must always be an obtuse angle.

4.4.3 The Convergence in Relative Entropy

We prove now a useful theorem which shows that convergence in relative entropy ($D(P||Q)$) implies convergence in the L_1 norm ($\|\star\|_1$) [45].

Definition 4.39 (L_1 norm)**.** [45] *The L_1 distance between two distributions P_1 and P_2 is given by: $\|P_1 - P_2\|_1 = \sum_{a\in X}|P_1(a) - P_2(a)|$.*

Lemma 4.40. *Let P_1 and P_2 be two probability distributions and let A be the set on which $P_1(x) > P_2(x)$, and let us suppose that $B \subseteq X$. Then:*
$max_{B\subseteq X}\{P_1(B) - P_2(B)\} = P_1(A) - P_2(A) = \dfrac{\|P_1 - P_2\|_1}{2}$.

Proof: (Due to [45]).

$$\begin{aligned}
\|P_1 - P_2\|_1 &= \sum_{x\in X} |P_1(x) - P_2(x)|, \\
&= \sum_{x\in A} (P_1(x) - P_2(x)) + \sum_{x\in A^c} (P_2(x) - P_1(x)), \\
&= P_1(A) - P_2(A) + P_2(A^c) - P_1(A^c), \\
&= P_1(A) - P_2(A) + (1 - P_2(A)) - (1 - P_1(A)), \\
&= 2(P_1(A) - P_2(A)), \\
\frac{\|P_1 - P_2\|_1}{2} &= P_1(A) - P_2(A), \\
max_{B\subseteq X}\{P_1(B) - P_2(B)\} &= P_1(A) - P_2(A) = \frac{\|P_1 - P_2\|_1}{2}. \ \square
\end{aligned}$$

Theorem 4.41. *The convergence in relative entropy implies convergence in the L_1 norm.*

Proof: (Also due to [45]). Consider two binary[23] probability distributions $P_1 = \{p, (1-p)\}$ and $P_2 = \{q, (1-q)\}$, and, without loss of generality, let us suppose that $p \geq q$. We need to show that:

$$\begin{aligned}
D(P_1||P_2) &\overset{a}{\geq} \frac{4}{2}(p-q)^2, \\
p\log\frac{p}{q} + (1-p)\log\frac{(1-p)}{(1-q)} &\overset{b}{\geq} \frac{4}{2}(p-q)^2, \\
g(p,q) &\overset{c}{\equiv} p\log\frac{p}{q} + (1-p)\log\frac{(1-p)}{(1-q)} - \frac{4}{2}(p-q)^2 \geq 0, \\
\frac{dg(p,q)}{dq} &= -\frac{p}{q} + \frac{(1-p)}{(1-q)} - \frac{4}{2}2(q-p), \\
\frac{dg(p,q)}{dq} &= \frac{(q-p)}{q(1-q)} - 4(q-p) \leq 0, \\
D(P_1||P_2) &\geq \frac{4}{2}(P_1(A) - P_2(A))^2, \\
D(P_1||P_2) &\overset{d}{\geq} \frac{1}{2}\|P_1 - P_2\|_1^2. \ \square
\end{aligned}$$

In a, we forget wittingly the conversion factor $\ln 2$ needed, in order to easily perform the derivative. The strange factor $\frac{4}{2}$ is used in step e. In b, we just substitute the binary distributions. In c, we define a function $g(p,q) \geq 0$, for $q \leq p$. Since $q(1-q) \leq \frac{1}{4}$, $\frac{1}{q(1-q)} \geq 4$. The inequality d is obtained squaring the result given in the Lemma 4.40. It shows that the convergence in relative entropy implies convergence in the L_1 norm.

[23] The proof is given just for the binary case.

4.5 The Second Law of Thermodynamics for Markovian Processes

Only entropy comes easy.

Anton Chekhov–Russian playwright and master of the modern short story, 1860-1904.

The entropy of an isolated system, in the micro-canonical ensemble, is defined as the logarithm of the number of the micro-states occupied, or in other words, $S = k_B \ln \Omega$ where the k_B is the Boltzmann constant, presented here just for dimensional reasons. This definition agrees with our intuition that, in an isolated system at thermal equilibrium, all micro-states are equally likely. The basic idea of Classical Statistical Mechanics is that every microscopic state possible occurs in practice because all measured properties are actually averages from all micro-states [28].

But why does entropy increase? This question is difficult to be answered for complex systems and for non-equilibrium systems. But for systems in *quasi-equilibrium*, we can construct a model consisting of a Markov chain in which the transitions obey all the physical laws governing the system [45]. Implicit in this assumption is the notion of an overall state of the system and the fact that if we know the present state of the system, its future is independent of the past. In such a system, we can find some different interpretations of the second law. It may come as a shock to find that the *entropy*, for these kind of processes, *does not always increase*. But *relative entropy always decreases* [45].

The assumption of Markovianity may appear to be a little bit strong, and it might be for some systems in Classical Statistical Mechanics. In fact, the Markovianity assumption can be stated if a system reaches the thermodynamic equilibrium. In Classical Information Theory, the systems are described by probability vectors, and in Quantum Mechanics, they are described by their density matrices each one representing, respectively, their classical and quantum distribution of probabilities. In Chapter 5, we will study the quantum systems and we will try to understand their evolution in the space of the density matrices as an analogy to the evolution of the classical probability vectors inside the probability simplex.

We showed that if a distribution evolves in the probability simplex by a bistochastic map, it is equivalent to understand this process as a Markov chain. We also presented in Theorem 4.23 that the stationary distribution for bistochastic evolution is the uniform vector. We shall find, somewhat surprisingly, that the second law is only true for Markov processes in which the equilibrium distribution is the uniform over the finite state space [58]. The next theorems will try to capture the essence of the Second Law of Thermodynamics, for Markovian processes.

Theorem 4.42. *For Markovian processes, the relative entropy $D(P_n||P_{n'})$ decreases when n increases, i.e., $D(p(x_n)||p^{'}(x_n)) \geq D(p(x_{n+1})||p^{'}(x_{n+1}))$.*

Proof: (Due to [45]). Suppose that P_n and $P_n^{'}$ denotes probability distributions at time n and let P_{n+1} and $P_{n+1}^{'}$ be the corresponding distributions at time

$n+1$. Thus, $P(x_n,\, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$ and $P^{'}(x_n,\, x_{n+1}) = p^{'}(x_n)r(x_{n+1}|x_n)$, where $r(\star|\star)$ is the probability transition function for the Markov chain. We know by the Bayes' Rule, that $p(x_n,\, x_{n+1}) = p(x_n)p(x_n|x_{n+1})$, or $p'(x_n,\, x_{n+1}) = p'(x_{n+1})p'(x_{n+1}|x_n)$, then using the Theorem 4.34, (the chain rule for relative entropy, we have):

$$D(p(x_n,\, x_{n+1})||p^{'}(x_n,\, x_{n+1})) = D(p(x_n)||p^{'}(x_n)) + D(p(x_{n+1}|x_n)||p^{'}(x_{n+1}|x_n)),$$
$$D(p(x_n,\, x_{n+1})||p^{'}(x_n,\, x_{n+1})) = D(p(x_{n+1})||p^{'}(x_{n+1})) + D(p(x_n|x_{n+1})||p^{'}(x_n|x_{n+1})).$$

But $D(p(x_{n+1}|x_n)||p^{'}(x_{n+1}|x_n)) = 0$, because $p(x_{n+1}|x_n) = p^{'}(x_{n+1}|x_n) = r(x_{n+1}|x_n)$. Thus by using the non-negativity of D, we prove the fact that relative entropy decreases when n increases: $D(p(x_n)||p^{'}(x_n)) \geq D(p(x_{n+1})||p^{'}(x_{n+1}))$. □

Theorem 4.43. *If the relative entropy between a distribution P_n and a stationary uniform distribution ($\mu \equiv \frac{1}{n}$) decreases with n, i.e., $D(P_n||\mu) \geq D(P_{n+1}||\mu) \geq \cdots \geq \cdots \geq 0$, then we have a Markovian process.*

Proof: [45] This Theorem is obvious since we have already proved in Theorem 4.23 that $P_n \to \mu$ for Markovian processes, *i.e.,* the uniform distribution is the stationary vector when the probability transition matrix is a bistochastic matrix. This sequence is a monotonically, non-increasing and nonnegative sequence and it must therefore have a limit. Then we can say that the limit is zero, because we know that $D(\star||\star) \geq 0$, so $D(P_n||\mu) \geq D(P_{n+1}||\mu)$ for all n and $D(P_n||\mu) \to 0$ when $n \to \infty$. □

Theorem 4.44. *Entropy increases in Markovian processes if the stationary distribution is uniform.*

Proof: The decrease in relative entropy does not implicate in a increase of entropy, in general. A simple counterexample is provided by any Markov chain with a non-uniform stationary distribution [45]. But, if the stationary distribution is the uniform distribution, then the entropy increases. This is our case because these type of processes are given by bistochastic matrices and we have already discussed in Theorem 4.23 that the uniform distribution is the stationary vector when the probability transition matrix is a bistochastic matrix. Let us define $\mu \equiv \frac{1}{n}$, then:

$$D(P_n||\mu) = \log n - H(P_n),$$

We can easily perceive that a monotonic decrease in the amount of D implies a monotonic increase in the entropy. This is the explanation that lies closer to statistical thermodynamics, where all the micro-states are equally likely [45]. And since the minimum of $D(P_n||\mu)$ is zero, the maximum value for the entropy is $\log n$. □

Theorem 4.45. *Suppose that we have a Markovian process such that $\vec{\tilde{p}} = B\vec{p}$ where B is a bistochastic matrix. Then $H(\vec{\tilde{p}}) > H(\vec{p})$.*

Proof: Since $\vec{\tilde{p}} = B\vec{p}$, by Theorem 4.7, we can say that $\vec{\tilde{p}} \prec \vec{p}$. We already know that the Shannon's entropy is a Schur concave function, then it *inverts* the majorization ordering: $H(\vec{\tilde{p}}) > H(\vec{p})$. We could prove this theorem by exhibiting the j-th element of the $\vec{\tilde{p}}$ vector in terms of the bistochastic transition matrix and later perform some calculations with the definition of entropy. □

4.6 The Jaynes Principle and the *maxent* Principles

4.6.1 The Principle of Insufficient Reason and The Jaynes Principle

The problem of finding a probability distribution by the knowledge of a subset of its moments, *i.e.*, the problem of determination of the remaining probabilities in cases where little information is reachable, is as old as the theory of probability. The *Laplace's principle of insufficient reason* was the first attempt in that sense [49]. It states that, in the discrete case, if n probabilities are indistinguishable and if you do not have any further information, the least biased distribution, which infers the minimum correlation between the unknown variables assigns all the n events equally likely [57]. The concept of entropy supplants the arbitrariness of Laplace's principle, and it presents a reasonable way to modify it when there are constraints and symmetries in the considered system [57, 59]. The statistical inference based on Shannon's information theory, in which the entropy function is a measure of our ignorance, claims that the most unbiased representation of a state is a probability distribution which has the maximum value possible of the entropy (the problem *maxent*) subject the constraints given by the information known a*priori* [49]. This principle is known in Classical Statistical Mechanics as the Jaynes principle and it is a generalization of Laplace's principle [49, 60]. In the following sections, we will justify these *maxent* principles.

4.6.2 The *maxent* Principles

The *maxent* Principle

The maximum-entropy principle [49, 60] is at the heart of statistical mechanics and thermodynamics. It allows one to select a probability distribution that is optimally unbiased, while considering the knowledge that is available for the system [59]. This principle states that the best probability distribution is that *i*) is compatible with our state of knowledge, *ii*) maximizes the entropy function under the constraints given by *i*). Since the entropy is a measure of the lacking information about a system, any set of probabilities that does not maximize the entropy contains additional unknown assumptions. Thus, it does not correspond to an unbiased choice [59]. We showed in Theorem 4.16 that if all probabilities are equally likely for a system with n possible outcomes, then the maximum principle follows immediately. More precisely, if we only have the normalization constraint, *i.e.*, with the constraint $\sum_x P(x) = \sum_x \frac{1}{n} = 1$, we

can say:

$$argmin_{\{x\in X\}}\{D(P(X)||\frac{1}{n})\} = argmax_{\{x\in X\}}\{H(P(X))\}. \tag{4.1}$$

In other words, the distribution that lies closer to the uniform distribution can be also achieved by maximizing the entropy. To make this problem more likely and interesting, we can add some other linear constraints. In Classical Statistical Mechanics, we maximize the entropy with further restrictions and constraints beyond the normalization. Thus, the *maxent* principle can be used to solve these constrained problems.

But what do these constraints mean physically? The n-dimensional state of knowledge about an experiment can be described by a set of mean values of some physical quantities, namely $\langle A_1\rangle$, $\langle A_2\rangle$, $\cdots$, $\langle A_n\rangle$, for example, we may have n linear constraints for these mean values like[24] $\langle E\rangle = E_0$, $\Rightarrow \sum_i p_i E_i = E_0$. Recalling that in Section 3.2, we have discussed that we can understand completely a physical system first describing an algebra for the observables and the state of the physical system is fully represented by expectations of all observables over the space. These linear constraints might be the energy, the temperature of a fluid, the average number of particles, or the distribution of velocities of the particles in movement, etc.

Laplace's principle of insufficient reason assumes uniformity when we do not have enough information. But clearly this is not the point here anymore. Because of this amount of information given by these linear constraints, the uniform distribution is not the solution for the optimization problem given by the Eq. 4.1. With this fact in mind, we can define the *maxent* problem as an optimization problem as done in Eq. 4.2

$$max_{\{p_i\in X\}}\{H(\{p_i\}),\ \text{s.t.}\ \sum_i p_i = 1, \sum_i p_i E_i = E_0\}. \tag{4.2}$$

We can construct a Lagrangian function to solve this optimization problem referred to the Eq. 4.2, [61]. This functional is constructed with the help of the Lagrange multipliers, thus:

The *maxent* problem:

$$L = -\sum_i p_i \ln p_i + \lambda\left(\sum_i p_i - 1\right) + \beta\left(\sum_i p_i E_i - E_0\right). \tag{4.3}$$

As the entropy[25] function H is a continuous function of p_i, the solution for the problem 4.3, stated in form of a lagrangian function, is given by:

$$\frac{\partial L}{\partial p_i} = 0,\ \forall i \in X, \tag{4.4}$$

$$\frac{\partial L}{\partial \lambda} = 0, \tag{4.5}$$

$$\frac{\partial L}{\partial \beta} = 0. \tag{4.6}$$

[24] We will change notations here to simplify the calculations, then $x \to i$ in order to ensure that we are working with discrete problems, but of course this formalism can be extended to continuous basis.

[25] The factor k_B - the Boltzmann constant - is set equal to 1.

Performing these derivatives defined in Eqs. 4.4, 4.5 and 4.6, we find:

$$-\ln p_i - 1 + \lambda + \beta E_i = 0, \ \forall i \in X, \tag{4.7}$$

$$\sum_i p_i - 1 = 0, \tag{4.8}$$

$$\sum_i p_i E_i - E_0 = 0. \tag{4.9}$$

Solving these equations 4.7, 4.8 and 4.9, we obtain the probability distribution that fulfill these equations:

$$p_i = e^{(1-\lambda)} e^{(-\beta E_i)}. \tag{4.10}$$

The Lagrange multiplier λ exists just for normalization reasons, and it is settled in order for the probability distribution to be normalized, *i.e.*, $\sum_i p_i = 1$. Thus we finally get the equilibrium distribution:

The solution of the *maxent* problem:

$$P^* \equiv \{p_i\}_{i=1}^n, \text{ such that } p_i = \frac{e^{-\beta E_i}}{Z}. \tag{4.11}$$

Where $Z \equiv \sum_i e^{-\beta E_i}$ is the partition function. This function plays an important role in Mechanical Statistics, see, for example, [28, 52]. The functional[26] L can be re-written by substituting the solution, *i.e.*, the equilibrium probability distribution p_i (Eq. 4.11), and having in mind that $\ln p_i = -\ln Z - \beta E_i$:

$$L = -\sum_i p_i[-\ln Z - \beta E_i] - \beta[\sum_i p_i E_i - E_0],$$

$$L = \sum_i p_i \ln Z + \beta \sum_i p_i E_i,$$

$$L = \ln Z + \beta E_0.$$

The dependence on the Lagrange multiplier λ vanishes because the corresponding constraint is fulfilled [59]. Then the equation $\frac{\partial L}{\partial \beta} = 0$ implies that $E_0 = -\frac{\partial \ln Z}{\partial \beta}$ in equilibrium. And as said before, when all constraints are fulfilled, we have $L = H$, then $H = \ln Z + \beta E_0$.

Physical Meaning of the Lagrange Multipliers

Let us consider a system[27] composed by two isolated and independent subsystems 1 and 2. When the two subsystems are brought into contact, the values of the energy may change individually. But as these two subsystems are isolated, the total energy must remain constant, $E_{tot} = E_1 + E_2 = const$. The total entropy[28] of the system is additive. Let us define a slack variable or a

[26] This functional L is identical to the entropy function because, when all constraints are fulfilled, the constraints will vanish.

[27] This subsection is totally inspired by [59].

[28] Here we use the notation S for entropy instead H because of the connection with Thermodynamics, but, in this text, we prefer to write H for the Shannon-Gibbs entropy and S for the Von Neumann entropy.

free parameter λ, which is adjusted via transitions between states in order to maximize the entropy, hence $\lambda \equiv E_1 - E_2$, then $E_1 = \frac{E_{tot}}{2} + \frac{\lambda}{2}$ and $E_2 = \frac{E_{tot}}{2} - \frac{\lambda}{2}$, $S(E_{tot}, \lambda) = S_1(E_1) + S_2(E_2) = S_1(\frac{E_{tot}}{2} + \frac{\lambda}{2}) + S_2(\frac{E_{tot}}{2} - \frac{\lambda}{2})$. Then, by using the *maxent* principle, we have:

$$\frac{\partial}{\partial \lambda} S(E_{tot}, \lambda) = 0,$$
$$\frac{\partial}{\partial \lambda} S(E_{tot}, \lambda) = \frac{1}{2}\frac{dS_1}{dE_1} - \frac{1}{2}\frac{dS_2}{dE_2} = 0,$$
$$\frac{dS_1}{dE_1} = \frac{dS_2}{dE_2}.$$

Hence the energy moves back and forth between the two subsystems until the system reaches a point, where the increase of the entropy of one system is identical to the loss of entropy of the other [59].

Now we must show that the derivative of the entropy is related to a Lagrange multiplier. We start from this equation $S = H = \ln Z + \beta E_0$, then:

$$\frac{\partial S}{\partial E_0} = \frac{d}{d\beta} \ln Z(\beta) \frac{d\beta}{dE_0} + \frac{d\beta}{dE_0} E_0 + \beta,$$
$$\frac{\partial S}{\partial E_0} = \left[\frac{d}{d\beta} \ln Z(\beta) + E_0\right] \frac{d\beta}{dE_0} + \beta.$$

We know that $E_0 = -\frac{d}{d\beta} \ln Z(\beta)$, then $[\frac{d}{d\beta} \ln Z(\beta) + E_0] = 0$, hence, we will find the expected result: $\frac{\partial S}{\partial E_0} = \beta$.

Theorem 4.46. *Let $Q = \{q_i\}_{i=1}^n$ be a probability distribution that satisfies all the constraints of the problem* maxent *(Eq. 4.3), and let $P^* = \{p_i\}_{i=1}^n$ be the solution of the same problem (Eq. 4.11). Then $H(Q) \leq H(P^*)$, and the equality holds iff $Q = P^*$.*

Proof: Let Q be a distribution that satisfies all the constraints of the *maxent* problem, $\sum_i p_i = \sum_i q_i = 1$ and $\sum_i p_i E_i = \sum_i q_i E_i = E_0$, then:

$$H(Q) = -\sum_i q_i \ln q_i,$$
$$H(Q) = -\sum_i q_i \ln\left(\frac{q_i p_i}{p_i}\right),$$
$$H(Q) = -\sum_i q_i \ln\left(\frac{q_i}{p_i}\right) - \sum_i q_i \ln p_i,$$
$$H(Q) = -D(Q||P^*) - \sum_i q_i \ln p_i,$$
$$H(Q) \overset{a}{\leq} -\sum_i q_i \ln p_i,$$
$$H(Q) \overset{b}{\leq} -\sum_i q_i \ln\left(\frac{e^{-\beta E_i}}{Z}\right),$$
$$H(Q) \leq -\sum_i q_i(-\beta E_i - \ln Z),$$

$$H(Q) \leq \beta \sum_i q_i E_i + \sum_i q_i \ln Z,$$
$$H(Q) \overset{c}{\leq} \beta E_0 + \ln Z = H(P^*),$$
$$H(Q) \leq H(P^*). \square$$

The inequality a is due to the positivity of D. In b, we only substituted the value of the equilibrium distribution $P^* = \{p_i\}$ given in Eq. 4.11 and in c, we use the fact that Q also satisfies all the constraints of the problem.

Corollary 4.47. *$H(P^*)$ is the global maximum point of the* maxent *problem.*

Proof: This proof is straightforward since we have already shown that the *maxent* problem is a concave problem over a compact and convex set. We have shown that $H(Q) \leq H(P^*)$, for all distributions Q which satisfies the constraints, and we also know that the concave problems must have only one point of maximum, then the global maximum point is reached when $Q = P^*$ with $H(Q) = H(P^*)$.

4.6.3 The *minRelativeEntropy* Principle

This problem (minimize the relative entropy), is a convex problem over a convex set and it has a global minimum. The Theorem 4.36 relates the unconstrained *maxent* problem with also the unconstrained *minRelativeEntropy* problem, but it is not the case here. We do not have any reason to adopt the uniform distribution as the stationary distribution. Let the stationary distribution be called P_0. If our process is described by a Markovian process, this distribution P_0 might be equal to the uniform distribution. Analogously, we can follow the same steps for the *maxent* problem, then:

$$min_{\{p_i \in X\}} \{D(P||P_0), \text{ s.t. } \sum_i p_i = 1, \sum_i p_i E_i = E_0\}. \tag{4.12}$$

We can construct a Lagrangian function to solve the problem 4.12. This function is constructed also with the help of the Lagrange multipliers, then:

The *minRelativeEntropy* problem:

$$L = \sum_i p_i \ln\left(\frac{p_i}{p_{0i}}\right) + \lambda\left(\sum_i p_i - 1\right) + \beta\left(\sum_i p_i E_i - E_0\right). \tag{4.13}$$

We find the solution for the problem 4.13 doing the same derivatives given by 4.4, 4.5 and 4.6, then find another equilibrium distribution:

The solution of the *minRelativeEntropy* problem:

$$p_i = p_{0i}\frac{e^{-\beta E_i}}{Z}, \tag{4.14}$$
$$Z = \sum_i p_{0i} e^{-\beta E_i}. \tag{4.15}$$

4.7 Some Other Classical Statistical Inference Schemes

> As a rule, probable inference is more like measuring smoke than counting children...
>
> Richard Cox

In the following subsections, we will give some examples of some classical statistical estimation schemes.

4.7.1 The Classical Maximum Likelihood Principle

Firstly, R. A Fisher published an interesting prototype of a numerical procedure in 1912 in [62], and, ten years later, the maximum likelihood principle was introduced formally by him in this article: [63], which considers Fisher's changing justifications for the method, the concepts he developed around it (including likelihood) and all approaches he discarded including, for example, inverse of probability [64].

Consider a family of probability distributions on $\mathcal{R}^n$ indexed by a vector $\vec{x}$, with densities $p_x(\star)$. When considered as a function of $\vec{x} = (x_1, x_2, \cdots, x_n)^t$ for some vector $\vec{y}$, the function $p_x(\vec{y})$ is called the likelihood function ($\mathcal{L}$) [61]. It is convenient to work with its logarithm $\log(\mathcal{L})$, the log-likelihood function. Since logarithm is a strictly monotonic and increasing function of $\vec{y}$, it is obvious that maximizing $\mathcal{L}$ is the same of maximizing the functional $\log \mathcal{L}$.

Definition 4.48 (The Likelihood function)**.** $\mathcal{L} = p_{\vec{x}}(\vec{y})$.

Definition 4.49 (The Log-Likelihood function)**.** $\log \mathcal{L} = \log p_{\vec{x}}(\vec{y})$.

There are often constraints on the values of the vector $\vec{x}$ which can represent prior knowledge about $\vec{x}$, or the domain of the likelihood function. Now, consider the problem of estimating the value of the $\vec{x}$ based on observing one sample y from the distribution [61]. A commonly used method in order to estimate $\vec{x}$ is called Maximum Likelihood (ML) estimation and it is shown in Eq. 4.16 [61].

$$\vec{x}_{ML} = argmax_{\{\vec{x} \in C\}} \{\log \mathcal{L} = \log p_{\vec{x}}(\vec{y})\}. \tag{4.16}$$

The Eq. 4.16 says that we need to choose, as our estimate, a value of the parameter that maximizes the log-likelihood function for the value of $\vec{y}$ that we observed [61]. Therefore the problem of finding a maximum likelihood estimate of the vector $\vec{x}$ can be stated as:

$$max\{\log \mathcal{L} = \log p_{\vec{x}}(\vec{y})\}. \tag{4.17}$$

$$\vec{x} \in C. \tag{4.18}$$

where C is the restriction set, nearly always a convex set. This constraint gives the prior information on the parameter vector $\vec{x}$, or in other words $\vec{x} \in C$. The maximum likelihood estimation problem Eq. 4.17 is a convex optimization problem if the log-likelihood function $\mathcal{L}$ is concave for each value of y_i, and

the set C can be described by a set of linear equality and convex inequality constraints, a situation which occurs in many estimation problems.

Let us reproduce a good example given by Fisher, in 1922, in this article: [63]. "A certain proportion, p of an infinite population is supposed to be a certain kind *e.g., successes*, the remainder are then *failures*. A sample of size equal to n is taken and found to contain a number x of successes and y failures, then $x + y = n$. The chance of obtaining such a sample is [63]":

$$\frac{n!}{x!\,y!}p^x(1-p)^y.$$

Applying the method of maximum likelihood, we have:

$$\log \mathcal{L} = \log\left(\frac{n!}{x!\,y!}\right) + x\log p + y\log(1-p),$$

Whence, by differentiating with respect to the parameter p, and obtaining the maximum value of the log-likelihood function and substituting $x + y = n$, we find the expected solution:

$$\frac{x}{p} = \frac{y}{1-p},$$
$$p = \frac{x}{n}.$$

4.7.2 The Classical Maximum Likelihood Principle and the Parametric Model

Suppose there is a set $x_1, x_2, \cdots, x_n$ of n independent and identically distributed observations, (IID) coming from an unknown distribution p_0. But it is supposed that this unknown distribution belongs to a certain family of distributions, called parametric model $\{p(\star|\theta),\ \theta \in C\}$, so that $p_0 = p(\star|\theta_0)$. The *true* value of the parameter θ_0 is assumed as unknown. We would like to use the Maximum Likelihood inference scheme in order to reach a value, as close as possible, to the true value of θ_0. Regarding to use of the Maximum Likelihood method, we must first specify the joint density function for the observed data. For an IID sample, this function is given by Eq. 4.19 [61].

$$p(x_1, x_2, \cdots, x_n|\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta). \tag{4.19}$$

The vector $\vec{x} = (x_1, x_2, \cdots, x_n)$ is considered to be a fixed parameter and θ is the variable allowed to vary freely. Then we can construct a Likelihood function:

$$\mathcal{L}(\theta|x_1, x_2, \cdots, x_n) = \prod_{j=1}^{n} p(x_j|\theta). \tag{4.20}$$

And the Log-likelihood can be defined as:

$$\log \mathcal{L}(\theta|x_1, x_2, \cdots, x_n) = \sum_{j=1}^{n} \log p(x_j|\theta). \tag{4.21}$$

Therefore we need to solve the following optimization problem:

$$max_{\{\theta\in\Theta\}}\{\log \mathcal{L}(\theta|x_1,\, x_2,\, \cdots,\, x_n) = \sum_{j=1}^{n} \log p(x_j|\theta).\},$$
$$\vec{x} \in C,$$
$$\theta \in \Theta.$$

For solving this problem and other similar problems based on ML estimate, see [61].

Linear Measurements with IID Noise

We consider a linear measurement model given by:

$$y_i = a_i^t x + \epsilon_i, \quad \forall i = 1,\, \cdots,\, m.$$

where $\vec{x} \in \mathcal{R}^n$ is a vector of parameters to be estimated, $y_i \in \mathcal{R}$ are the measured quantities and the ϵ_i are the IID noise or error [61]. The likelihood function is then:

Definition 4.50 (The Likelihood function). $\mathcal{L} = \prod_{i=1}^{m} p(y_i - a_i^t \vec{x})$.

And the log-likelihood function is:

Definition 4.51 (The Log-Likelihood function). $\log \mathcal{L} = \sum_{i=1}^{m} \log p(y_i - a_i^t \vec{x})$.

The ML estimate is any optimal point for the optimization problem:

$$max\{\log \mathcal{L} = \sum_{i=1}^{m} \log p(y_i - a_i^t \vec{x})\}, \tag{4.22}$$
$$\vec{x} \in C. \tag{4.23}$$

Maximum a *posteriori* probability estimation - MAP

Maximum a *posteriori* probability (MAP) estimation can be considered as a Bayesian version of the maximum likelihood estimation, with a prior probability density on the underlying parameter $\vec{x}$ [61]. We assume that $\vec{x}$ (the vector to be estimated) and $\vec{y}$ (the observation) are random variables with a joint probability density given by $p(x,\, y)$. The prior density of x is given by:

$$p_x(x) = \sum_y p(x,\, y).$$

This density represents our prior information on which values of the vector x might be, before we observe the vector $\vec{y}$. Similarly, the prior density of y is given by [61]:

$$p_y(y) = \sum_x p(x,\, y).$$

This density represents the prior information on which measurements or observations of vector $\vec{y}$ will be. The conditional density function of $\vec{y}$, given $\vec{x}$, is given by:

$$p(y|x) = \frac{p(x, y)}{p_x(x)}.$$

In the MAP estimation method, $p(y|x)$ plays the role of the parameter dependent density p_x in the maximum likelihood estimation setup. The conditional density of $\vec{x}$, given $\vec{y}$, is given by:

$$p(x|y) = \frac{p(x, y)}{p_y(y)} = p(y|x)\frac{p_x(x)}{p_y(y)}.$$

In the MAP estimation method, our estimate of $\vec{x}$, given the observation $\vec{y}$, is given by:

$$\vec{x}_{MAP} = argmax_{\{\{x_i\}\in C\}}\{\mathcal{L}_{MAP} = p(x, y)\}, \tag{4.24}$$

$$\vec{x}_{MAP} = argmax_{\{\{x_i\}\in C\}}\{\mathcal{L}_{MAP} = p(y|x)p_x(x)\}. \tag{4.25}$$

Then, we take, as estimate of $\vec{x}$, the value that maximizes the conditional density of $\vec{x}$, given the observed value of $\vec{y}$ [61]. The only difference between this estimate and the maximum likelihood estimate is the second term, $p_x(x)$, appearing here. This term can be interpreted as taking our prior knowledge of $\vec{x}$ into account [61]. Taking logarithms:

$$\vec{x}_{MAP} = argmax_{\{\{x_i\}\in C\}}\{\log \mathcal{L}_{MAP} = \log p(y|x) + \log p_x(x)\}. \tag{4.26}$$

The first term is essentially the same as the log-likelihood function; the second term penalizes choices of $\vec{x}$ that are unlikely, according to the prior density [61]. The only difference between the ML estimate and the MAP estimate is the presence of an extra term in the optimization problem, associated with the prior density of $\vec{x}$. [61].

CHAPTER 5

Introduction to Quantum Information Theory

> Anything you can do in classical physics, we can do better in quantum physics.
>
> Daniel Kleppner

We will try to establish an *invisible* parallel with the Classical Information theory, developed in the Chapter 4. Then it would be nice the reader had in mind all previously studied sections.

5.1 Operator Majorization and Partial Ordering

The connection of the Theorem 4.8 with the Quantum Mechanics theory is due to the notion of the majorization of operators. We can define n-dimensional operators, such as $\rho \prec \sigma$, if we define vectors $\lambda(\rho) \prec \lambda(\sigma)$ whose components are the eigenvalues of the operators arranged in decreasing order [38]. The Uhlmann's theorem is the operator analogue to the Theorem 4.8. It allows us to quantify the randomness of a hermitian operator (eventually a density matrix).

Theorem 5.1 (Uhlmann's Theorem)**.** *Let ρ and σ be two hermitian matrices. Then $\rho \prec \sigma$ iff there exists unitary matrices U_i and a probability distribution $\vec{p} = \{p_i\}$ such that $\rho = \sum_{i=1}^{n} p_i U_i \sigma U_i^\dagger$.*

Proof: (Just the commutative case). The general proof for this theorem is out of the scope of this text. It is obvious that Theorem 5.1 is the operator analogue to Theorem 4.8, even if the matrices do not commute. However, for the commutative case, this fact is easy to be shown, since there already exists

a basis where these two theorems are the same. We can always diagonalize a hermitian matrix, and we can always find a basis where the two commutative matrices are diagonal. Then we can define diagonal vectors and, by using Theorem 4.8, we can see that $diag(\vec{\lambda}(\rho)) = \sum_{i=1}^{n} p_i \Pi_i diag(\vec{\lambda}(\sigma))$ or in other words, there exists a bistochastic matrix B such as $diag(\vec{\lambda}(\rho)) = B\ diag(\vec{\lambda}(\sigma))$ which implies that $diag(\vec{\lambda}(\rho)) \prec diag(\vec{\lambda}(\sigma))$. □

5.2 Some Other Results in Operator Majorization Theory

5.2.1 Stochastic and Bistochastic Maps

Lemma 5.2 (Quantum HLP Lemma)**.** *There exists a completely positive bistochastic map transforming ρ into σ iff $\rho \prec \sigma$.*

Proof: (Given by [31], just the *only if*). Let us introduce unitaries U and V such that $diag(\vec{\lambda}(\sigma)) = U\sigma U^{\dagger}$ and $diag(\vec{\lambda}(\rho)) = V\rho V^{\dagger}$. Given a bistochastic map such that $\Phi(\rho) = \sigma$, we can always construct another bistochastic map Ψ such that $\Psi(\star) = U\{\Phi[V^{\dagger}(\star)V]\}U^{\dagger}$. We know that by hypothesis $\Psi[diag(\vec{\lambda}(\rho))] = diag(\vec{\lambda}(\sigma))$. Let us define a matrix by its matrix elements as $B_{ij} \equiv Tr(P_i \Psi P_j)$, where the P_i are projectors and B is a bistochastic matrix. Finally $\lambda_i(\sigma) = Tr[P_i diag(\vec{\lambda}(\rho))] = Tr[P_i \Psi(\sum_j P_j \lambda_j(\rho))] = Tr[\sum_j (P_i \Psi P_j)\lambda_j(\rho))] = \sum_j B_{ij}\lambda_j(\rho)$. An appeal to the classical Theorem HLP (Theorem 4.7) concludes the proof [31].

Theorem 5.3 (Schur-Horn Theorem)**.** *Let ρ be a hermitian matrix, $\vec{\lambda}$ its spectrum, and $\vec{p}$ its diagonal elements in a given basis. Then $\vec{p} \prec \vec{\lambda}$.*

Proof: (Given by [31]) Let us suppose that the unitary U diagonalizes the matrix, then $p_i = (\rho)_{ii} = \sum_{jk} U_{ij}\lambda_j \delta_{jk} U^{\dagger}_{ki} = \sum_j |U_{ij}|^2 \lambda_j$. These two vectors are linked by an unistochastic matrix, hence a bistochastic matrix, then $\vec{p} = B\vec{\lambda}$ or $\vec{p} \prec \vec{\lambda}$.

Corollary 5.4. *If $\vec{p} \prec \vec{\lambda}$, then there exists a hermitian matrix with spectrum $\vec{\lambda}$ whose diagonal elements are given by the entries of $\vec{p}$ in a given basis.*

Proof: The proof is the reverse of the proof of the Theorem 5.3, and this matrix is given by B.

Pre-order in the Space of Bistochastic Matrices

We showed in Theorem 4.4 that the set of the n-dimensional bistochastic matrices is the convex hull of the set of $n!$ permutation matrices of order n [31]. To settle which bistochastic matrix have stronger mixing properties, one may introduces a relation (pre-ordering) in the space of bistochastic matrices [44]. Let B_1 and B_2 be two bistochastic matrices. Then we have the following property: $B_1 \prec B_2$ *iff* there exists[1] another bistochastic matrix B such that $B_1 = BB_2$ [44]. We need to distinguish some bistochastic matrices: the Waerden matrix B_* which has all its n^2 elements equal to $\frac{1}{n}$ and the permutation matrices Π. For any arbitrary bistochastic matrix B we have

[1] The proof will not be discussed here and it can be seen in [65].

$B_* = BB_*$ and $B = (B\Pi^{-1})\Pi$, and hence $B_* \prec B \prec \Pi$ [44]. The relation $B_1 \prec B_2$ implies that $B_1\vec{x} \prec B_2\vec{x}$, for any $\vec{x} \in \mathcal{R}_+^n$ [44].

Theorem 5.5. *If P and P^{-1} (its inverse) are both bistochastic matrices, then P and P^{-1} are permutation matrices.*

Proof: Hypothesis: $PP^{-1} = \mathbb{I}$, and both matrices are bistochastic. Then $P\vec{x} \prec \vec{x}$ for all $\vec{x} \in \mathcal{R}_+^n$, but $\vec{x} = (P^{-1}P)\vec{x}$, then $\vec{x} \prec P\vec{x}$, for all $\vec{x} \in \mathcal{R}_+^n$. Hence $\vec{x} \prec P\vec{x} \prec \vec{x}$ for all $\vec{x} \in \mathcal{R}_+^n$. This fact allows us to say that P is a permutation matrix, because it majorizes a vector, but it is always majorized by the same vector, (and this happened because both matrices are bistochastic and one is the inverse of the another). Then it is obvious that this matrix P just changes the position of the elements of the vectors. We could do the same thing using P^{-1} instead of P, then both matrices are permutation matrices. □

Other Interesting Results

Theorem 5.6 (Ky Fan's Principle)**.** *For any hermitian matrix A, the sum of the k largest eigenvalues of A is the maximum value of $Tr(AP)$, where the maximum is taken over all k-dimensional projectors P, or: $\sum_{i=1}^{k} \lambda_j(A) = max_{dim(P)=k}\{Tr(AP)\}$.*

Proof: Let us define the projector P as the projector onto the k dimensional subspace spanned by the k eigenvectors of A associated with the k largest eigenvalues. Then $Tr(AP) = \sum_{i=1}^{k} \lambda_k(A)$. We just need to prove that $Tr(AP) \leq \sum_{i=1}^{k} \lambda_k(A)$ for any k dimensional projector P. Let $\{|e_i\rangle\}_{i=1}^{n}$ be an orthonormal basis $P = \sum_{i=1}^{k} |e_i\rangle\langle e_i|$ [38]. And let $\{|f_i\rangle\}_{i=1}^{n}$ be a set such that $A|f_i\rangle = \lambda_i|f_i\rangle$. Then $\vec{a} = \langle e_i|A|e_i\rangle = \sum_{i=k}^{n} |u_{ik}|^2\lambda_k$, where the numbers $|u_{ik}|^2 = \langle e_i|f_k\rangle\langle f_k|e_i\rangle$. The matrix defined by its matrix elements (u_{ik}) is bistochastic. Then $\vec{a} \prec \vec{\lambda}(A)$. Thus $\sum_{i=1}^{k} a_i = \sum_{i=1}^{k} \langle e_i|A|e_i\rangle \leq \sum_{i=1}^{k} \lambda_i(A)$ [38]. □

Corollary 5.7. *The Ky Fan's principle implies that for hermitian matrices A and B, $\vec{\lambda}(A+B) \prec \vec{\lambda}(A) + \vec{\lambda}(B)$.*

Proof: To prove this corollary, choose the k dimensional projector as $P = \sum_{i=1}^{k} \lambda_i(A+B) = Tr[(A+B)P] = Tr(AP) + Tr(BP) \leq \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B)$ [38].

Theorem 5.8. *Let ρ be a density matrix, $\vec{p} = \{p_j\}$ a probability distribution and ρ_j density matrices such that $\rho = \sum_i p_i\rho_i$. Then $\vec{\lambda}(\rho) \prec \sum_i p_i\vec{\lambda}(\rho_i)$.*

Proof: The proof follows immediately from the Corollary 5.7 [38]. □

5.3 Quantum Maps and Quantum Operations

5.3.1 Positive and Completely Positive Maps

Let us suppose that we want to describe the evolution of the quantum systems, so the *quantum operation* formalism[2] is the correct tool for describing various

[2] This section is just informative and we will not spend a lot of time on this issue. A wonderful approach of this theme can be seen in [27].

processes, including stochastic changes to quantum states, in a similar manner of the stochastic and bistochastic processes which are described by the Markov chains in the classical world of probabilities [10]. Just as in the classical case, where the state is described by a vector of probabilities, we shall describe the quantum states in terms of the density operator ρ. We can also describe a classical system by its density operator (see [23, 25]). So, the classical states are transformed by this rule: $\vec{p}' = S\vec{p}$, and the quantum states are transformed as:

$$\rho' = \mathcal{E}(\rho). \tag{5.1}$$

All unitary transformations can be written as $\mathcal{E}(\rho) = U\rho U^\dagger$, and measurements can also be written as $\mathcal{E}(\rho) = M_m \rho M_m^\dagger$, using the notation given by the Eq. 5.1. There exists *two* separate ways of understanding quantum operations, (both ways are equivalent) [10]:

1. Environmental representation.
2. Kraus evolution.

1. Environmental Representation

The environmental representation is the natural way to describe the dynamics of a quantum open system. The system is divided in two parts, or subsystems: the principal subsystem and the environment, which together form a closed quantum system [10, 31]. The principal subsystem is allowed to freely interact with the environment. Let us assume that the composite system is a product state: $\rho = \rho_S \otimes \rho_E$. Then we apply a global unitary $U_{SE} \in \mathcal{H}_S \otimes \mathcal{H}_E$ in the system, (it describes the interaction of the two subsystems) and then we perform a *partial* trace[3] on the environment. This operation allows us to eliminate the environment and keep the only subsystem that interests us. This process is analogue to the classical process of separation of a subsystem from a system.

Then $\mathcal{E}(\rho) = Tr_E[U_{SE}(\rho_S \otimes \rho_E)U_{SE}^\dagger] = \rho'_S$. That is, the whole process can be described as $\mathcal{E}(\rho) = \rho'_S$: $\rho \to (\rho_S \otimes \rho_E)$, then $(\rho_S \otimes \rho_E) \to U_{SE}(\rho_S \otimes \rho_E)U_{SE}^\dagger$, and then $Tr_E[U_{SE}(\rho_S \otimes \rho_E)U_{SE}^\dagger] = \rho'_S$.

We initially assume that the two subsystems can be written in a product state. But of course this is not true in general, because if two quantum systems are interacting, some correlation may appear. However this assumption, generally, is quite reasonable. This method has a great disadvantage: it is difficult to specify the d^2 global unitary that acts on the system [10].

2. Kraus Evolution

A linear map Φ is completely positive *iff* it can be written in its standard Kraus form [31]. Then a completely positive map $\Phi : \mathcal{M}_n \longmapsto \mathcal{M}_n$ can always be

[3] The partial trace is defined as $Tr_A(\rho_{AB}) = \rho_B = [Tr(\star)_A \otimes \mathbb{I}_B(\rho_{AB})]$, if $\{|k\rangle\}_{i=1}^{\dim(A)}$, is a basis for the space of the subsystem A, hence $\rho_B = [\sum_k |k\rangle\langle k| \otimes \mathbb{I}](\rho_{AB})$. Thus, summing the terms properly, $\rho_B = \sum_k[(\langle k| \otimes \mathbb{I})\rho_{AB}(|k\rangle \otimes \mathbb{I})]$. Therefore, $\rho_B = \sum_K \langle k|\rho_{AB}|k\rangle$.

represented as:

$$\rho \to \rho' = \sum_{i=1}^{r\leq n^2} d_i \chi_i \rho \chi_i^\dagger = \sum_{i=1}^{r} A_i \rho A_i^\dagger,$$

$$\text{where } Tr(A_i^\dagger A_j) = \sqrt{d_i d_j}\langle \chi_i | \chi_j \rangle = d_i \delta_{ij},$$

$$\text{if the map is trace-preserving, then: } \sum_{i=1}^{r} A_i^\dagger A_i = \mathbb{I}_n, \Rightarrow \sum_{i=1}^{r} d_i = n.$$

A connection with the environmental representation can be observed if we start the map by coupling an environmental ancilla $|\nu\rangle\langle\nu|$, described as a pure state, then: $\rho' = Tr_E[U_{SE}(\rho_S \otimes \rho_E)U_{SE}^\dagger]$, or $\rho' = Tr_E[U_{SE}(\rho_S \otimes |\nu\rangle\langle\nu|)U_{SE}^\dagger] = \sum_{\mu=1}^{k}[\langle\mu|U_{SE}|\nu\rangle\rho_S\langle\nu|U_{SE}^\dagger|\mu\rangle] = \sum_{\mu=1}^{k} A_\mu \rho_S A_\mu^\dagger$, obeying the completeness relation: $\sum_{\mu=1}^{k}[\langle\mu|U_{SE}^\dagger|\nu\rangle\langle\nu|U_{SE}|\mu\rangle] = \langle\nu|U_{SE}^\dagger U_{SE}|\nu\rangle = \mathbb{I}_n$ [31].

As an example let us start with the initial state of the environment as a maximally mixed state $\rho_E = \frac{\mathbb{I}}{n}$ [31]. The unitaries may be treated as vectors in a bigger space, the composite space $\mathcal{H}_{\mathcal{HS}A} \otimes \mathcal{H}_{\mathcal{HS}B}$ Hilbert-Schmidt space, and they are represented as $U = \sum_{i=1}^{n^2} \sqrt{\lambda_i}|\tilde{A}_i\rangle \otimes |\tilde{A'}_i\rangle$ [31]. The procedure of partial tracing leads us to a Kraus form with n^2 terms:

$$\rho \to \rho' = Tr_E[U_{SE}(\rho_S \otimes \frac{\mathbb{I}}{n})U_{SE}^\dagger],$$

$$= Tr_E[\sum_{i=1}^{n^2}\sum_{j=1}^{n^2}\sqrt{\lambda_i\lambda_j}(\tilde{A}_i\rho\tilde{A}_j{}^\dagger) \otimes (\frac{\mathbb{I}}{n}\tilde{A}'_i\tilde{A}'^\dagger_j)],$$

$$\rho' = \frac{1}{n}\sum_{i=1}^{n^2}\sum_{j=1}^{n^2}\lambda_i(\tilde{A}_i\rho\tilde{A}_i{}^\dagger),$$

the standard Kraus form is obtained if we define: $A_i \equiv \sqrt{\frac{\lambda_i}{n}}\tilde{A}_i$:

$$\rho' = \sum_{i=1}^{n^2}\sum_{j=1}^{n^2}(A_i\rho A_i^\dagger).$$

5.3.2 The Measurement Postulate and The POVM's

Selective Measurements

Let us define the space of all measurements outcomes consisting of k measurement operators M_i, with k possible outcomes, in which satisfies the completeness relation: $\sum_{i=1}^{r} M_i^\dagger M_i = \mathbb{I}_n$ [31]. As mentioned in Chapter 3, the quantum measurement performed on the initial state ρ produces the i-th outcome with probability p_i and it changes the initial state ρ into ρ_i:

$$\rho \to \rho_i = \frac{M_i\rho M_i^\dagger}{Tr(M_i\rho M_i^\dagger)},$$

$$\text{with probability: } p_i = Tr(M_i\rho M_i^\dagger).$$

All probabilities are positive and sum to 1. This measurement process is called *selective* measurement. If we do not post-select the state based on the k-th

outcome of the measurement, the initial state is transformed into a convex combination of all possible outcomes.

Projective Measurements

In a projective measurement process, the measurement operators are orthogonal projectors, so $M_i = P_i = P_i^\dagger = M_i^\dagger$, with $P_i P_j = P_i \delta_{ij}$ [31]. A projective measurement is defined by an observable, *i.e.*, a hermitian operator with spectral decomposition $\mathcal{O} = \sum_{i=1}^{n} \lambda_i P_i$, (let us suppose that all eigenvalues are non-degenerate). Using this decomposition, we obtain a set of orthogonal measurements operators, which forms a resolution of the identity operator (the eigenprojectors related to different eigenvalues in a hermitian operator are all orthogonal). All the possible outcomes are labeled by the eigenvalues of $\mathcal{O}$. If we do not post-select the state, it will be transformed into the following convex mixture:

$$\rho \to \rho' = \sum_{i=1}^{n} P_i \rho P_i, \tag{5.2}$$

$$[\rho', \mathcal{O}] = 0. \tag{5.3}$$

The i-th outcome (the i-th eigenvalue of $\mathcal{O}$), labeled as λ_i occurs with probability p_i, and the initial state is transformed into:

$$\rho \to \rho_i = \frac{P_i \rho P_i}{Tr(P_i \rho P_i)},$$
$$\text{with probability: } p_i = Tr(P_i \rho P_i) = Tr(\rho P_i).$$

As we have already discussed in Chapter 3, the expectation value of the observable is given by this rule: $\langle \mathcal{O} \rangle = Tr(\mathcal{O}\rho)$.

The POVM's

Definition 5.9 (Positive Operator Value Measures – POVM). *A POVM is defined as any partition of the identity operator into a set of k positive operators E_i that satisfies: $\sum_{i=1}^{k} E_i = \mathbb{I}$, with $E_i = E_i^\dagger$ and $E_i \geq 0$ for all $i = 1, \cdots, k$.*

As we may choose $E_i = M_i M_i^\dagger$, we can write POVM's in order to obtain selective and projective measurements. But they do not determine uniquely the measurement operators M_i, except in the special case of projective measurements [31]. A set with n^2 POVM's is called informationally complete if its statistics determine uniquely a density matrix. A POVM is said to be pure if each operator E_i is rank one. A set of k pure states defines a pure POVM *iff* the maximally mixed state $\rho^* = \frac{\mathbb{I}}{n}$ can be decomposed as $\rho^* = \sum_{i=1}^{k} p_i |\psi_i\rangle\langle\psi_i|$, with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$. For any set of POVM, just define the density matrices as $\rho_i \equiv \frac{E_i}{Tr(E_i)}$ and mix them in a convex mixture with weights defined as $p_i = Tr(\frac{E_i}{n})$, such that: $\sum_{i=1}^{k} p_i \rho_i = \sum_{i=1}^{k} \frac{E_i}{n} = \frac{\mathbb{I}}{n} = \rho^*$.

5.3.3 Unitary Transformations

Theorem 5.10 (Kadison's Theorem). *Let there be a map $\Phi : \mathcal{M}_n \longmapsto \mathcal{M}_n$ which is one-to-one and onto, and which preserves the convex structure in the sense that*

$\Phi(\sum_i p_i\rho_i) = \sum_i p_i\Phi(\rho_i)$, *must be the form* $\Phi(\rho) = U\rho U^\dagger$, *where* U *is a unitary matrix* [31].

The proof of the Theorem 5.10 is obviously out of the scope of this introductory text. In infinitesimal form, a unitary transformation takes the form:

$$\dot{\rho} = \frac{1}{i\hbar}[H, \rho]. \tag{5.4}$$

Where H is a hermitian operator, a hamiltonian, for example. The Eq. 5.4 is the analogue, in quantum mechanics, of the Liouville equation[4] and it describes the time evolution of a quantum mixed state. The procedure, often used to formulate some quantum analogies of classical systems, involves a description of classical systems by using the Hamiltonian mechanics. Classical variables are then interpreted as quantum operators, while Poisson brackets are replaced by quantum commutators, divided by $i\hbar$.

We know that the solution of Schrödinger's equation assuming a time-independent Hamiltonian is given by $|\psi(t)\rangle = U(t)|\psi(0)\rangle = e^{\frac{-iHt}{\hbar}}|\psi(0)\rangle$. The link between the Eq. 5.4 and the Theorem 5.10 can be noticed if we perform the solution of the differential equation involved [14]:

$$d\rho = -\tfrac{i}{\hbar}[H, \rho]dt,$$
$$\text{where } [H, \rho] = [H, \star](\rho).$$
$$\text{Thus, the solution of Eq. 5.4 is:}$$
$$\rho(t) = e^{-\frac{it}{\hbar}[H,\star]}\rho(0),$$
$$\rho(t) = \sum_{n=0}^{\infty}(-\frac{it}{\hbar})^n\frac{[H, \star]^n}{n!}\rho(0),$$
$$\rho(t) = \rho(0) - \frac{it}{\hbar}[H, \rho(0)] + \frac{1}{2}(\frac{-it}{\hbar})^2[H, [H, \rho(0)]] + \cdots,$$
$$\rho(t) = (\mathbb{I} - \frac{it}{\hbar}H + \frac{t^2}{2\hbar^2}H^2 - \cdots)\rho(0)(\mathbb{I} + \frac{it}{\hbar}H - \frac{t^2}{2\hbar^2}H^2 + \cdots),$$
$$\rho(t) = \left[\sum_{n=0}^{\infty}\frac{(\frac{-iHt}{\hbar})^n}{n!}\right]\rho(0)\left[\sum_{n=0}^{\infty}\frac{(\frac{iHt}{\hbar})^n}{n!}\right],$$
$$\rho(t) = e^{\frac{-iHt}{\hbar}}\rho(0)e^{\frac{iHt}{\hbar}},$$
$$\rho(t) = U(t)\rho(0)U(t)^\dagger.$$

Some Examples of One Qubit Quantum Gates

Let $\mathcal{H}_{2\times 2}$ be the one qubit space generated by the orthonormal basis $\{|0\rangle, |1\rangle\}$. Let us give some examples of unitary operators $U : \mathcal{H}_{2\times 2} \longmapsto \mathcal{H}_{2\times 2}$ [14]:

- $X = |0\rangle\langle 1| + |1\rangle\langle 0|$,

[4] When it is applied to the expectation value of an observable $\mathcal{O}$, the corresponding equation (Eq. 5.4) is given by the Ehrenfest's theorem [2], and takes the form $\frac{d}{dt}\langle\mathcal{O}\rangle = \frac{\langle[\mathcal{O},H]\rangle}{i\hbar} + \left\langle\frac{d\mathcal{O}}{dt}\right\rangle$. Obviously if the observable commutes with the Hamiltonian and it does not depend on time explicitly, then it is a constant of movement.

- $Y = -i|0\rangle\langle 1| + i|1\rangle\langle 0|$,
- $Z = |0\rangle\langle 0| - |1\rangle\langle 1|$,
- $H = |+x\rangle\langle 0| + |-x\rangle\langle 1|$, with $|\pm x\rangle = \frac{1}{\sqrt{2}}[|0\rangle \pm |1\rangle]$,
- Phase: $\Theta = |0\rangle\langle 0| + e^{i\theta}|1\rangle\langle 1|$.

5.4 The Von Neumann Entropy $S(X)$

Traditionally[5] entropy is derived from phenomenological and thermodynamic considerations based on the *second law of thermodynamics*. It relates the behavior of a macroscopic system in equilibrium or close to equilibrium. This interpretation, based on the second law, may frequently lead to obscure and very speculative or even mystical ideas[6] [66]. An accurate definition of entropy is only achievable in the realm of quantum mechanics, whereas in classical mechanics framework, entropy can only be introduced in a somewhat limited and artificial manner [66]. Confessedly entropy plays an important role among the physical quantities, although its nature is admittedly purely probabilistic. Therefore, a satisfactory and non-speculative interpretation for this quantity is only found in the quantum mechanics theory, that assigns it a measure of the amount of chaos within a quantum mixed state [66].

Entropy is quite different from most physical quantities. In quantum mechanics, we have a C^* non-commutative algebra for the observables and the states are defined by its expectation values in all observables. These states are described (as discussed in Chap 3) by density matrices ρ, *i.e.*, they are trace one positive hermitian operators and its expectation values are defined as $\langle \mathcal{O} \rangle = Tr(\mathcal{O}\rho)$. But entropy is not an observable, which means that there is not a hermitian operator, whose expected value in any state would be its entropy [66]. It is a function of a hermitian operator. Then if we want to define the entropy of a quantum state, we need first to understand what is a function of an operator:

Definition 5.11 (Operator Function)**.** *Let A be a hermitian operator. Then $A = U\Lambda U^\dagger$, for some unitary matrix U and $\Lambda = diag(\vec{\lambda}(A))$, is the diagonal matrix containing the eigenvalues of A. For any function $f : \mathcal{R} \longmapsto \mathcal{R}$, we can define $f(A) = U\, diag[f(\vec{\lambda}(A))]\, U^\dagger$.*

Theorem 5.12 (Klein's Inequality)**.** *If a function f is a convex function and A and B are hermitian operators, then $Tr[f(A) - f(B)] \geq Tr[(A - B)f'(B)]$, with equality iff $A = B$ [31].*

Proof: As A and B are hermitian operators, they can be diagonalized. Let us suppose that $A|e_i\rangle = a_i|e_i\rangle$ and $B|f_i\rangle = b_i|f_i\rangle$ such that $\langle e_i|f_j\rangle = c_{ij}$ [31].

[5] This section was totally inspired by the wonderful work of A. Wehrl ([66]).
[6] The heat death of the universe is a good example of these ideas.

Then:

$$\sum_i \langle e_i|f(A) - f(B) - (A-B)f'(B)|e_i\rangle,$$
$$\sum_i \langle e_i|f(a_i)|e_i\rangle\langle e_i| - f(b_j)|f_j\rangle\langle f_j| - (a_j - b_j)f'(b_j)|f_j\rangle\langle f_j||e_i\rangle,$$
$$= f(a_i) - \sum_j |c_{ij}|^2[f(b_j) - (a_i - b_j)f'(b_j)],$$
$$\stackrel{a}{=} \sum_j |c_{ij}|^2[f(a_i) - f(b_j) - (a_i - b_j)f'(b_j)] \stackrel{b}{\geq} 0. \square$$

The equality a holds because the summation is over j. The inequality b is because every differentiable convex function obeys $f(y) - f(x) \geq (y-x)f'(x)$.

Corollary 5.13 (Klein's Inequality - Special case)**.** *This corollary is a special case of the Theorem 5.12. If a function f is a convex function and A and B are hermitian operators, then: $Tr[A(\log A - \log B)] \geq Tr(A - B)$.*

Proof: Let us suppose, by hypothesis, that the inequality is true. Define a convex function f as: $f(\star) = (\star)\log(\star)$, with $f'(\star) = [\log(\star) + 1]$. After minor calculations, we find the result of the Theorem 5.12:

$$Tr[A(\log A - \log B)] \geq Tr(A - B),$$
$$Tr(A\log A) - Tr(A\log B) \geq Tr(A - B),$$
$$Tr(A\log A) - Tr(A\log B) - Tr(B\log B) \stackrel{a}{\geq} Tr(A - B) - Tr(B\log B),$$
$$Tr[A\log A - B\log B] \geq Tr(A - B) + Tr(A\log B) - Tr(B\log B),$$
$$Tr[A\log A - B\log B] \geq Tr(A - B) + Tr[(A - B)(\log B)],$$
$$Tr[A\log A - B\log B] \geq Tr[(A - B)(\log B + \mathbb{1})],$$
$$\text{But } f(\star) = (\star)\log(\star), \text{ and } f'(\star) = [\log(\star) + 1], \text{ then:}$$
$$Tr[f(A) - f(B)] \geq Tr[(A - B)f'(B)]. \square$$

In a, we just add the factor $[-Tr(B\log B)]$ in both sides of the inequality and, everywhere, we use the linearity of the trace function.

Theorem 5.14 (Peierl's Inequality)**.** *If a function f is a strictly convex function and A is a hermitian operator, then $Tr[f(A)] \geq \sum_i f(\langle f_i|A|f_i\rangle)$, where $\{|f_i\rangle\}$ is a complete set of orthonormal vectors and the equality holds iff $|f_i\rangle = |e_i\rangle$, where $A|e_i\rangle = a_i|e_i\rangle$* [31].

Proof: The function f is a convex function. Just observe that for any vector $|f_i\rangle$ we have: $\langle f_i|A|f_i\rangle = \sum_j |\langle f_j|e_i\rangle|^2 f(a_i) \geq f(\sum_j |\langle e_i|f_j\rangle|^2 a_j) = f(\langle f_i|A|f_i\rangle)$. Summing over i gives the result: $Tr[f(A)] = \sum_i \langle f_i|A|f_i\rangle = \sum_i \sum_j |\langle f_j|e_i\rangle|^2 f(a_i)$ $\geq \sum_i f(\sum_j |\langle e_i|f_j\rangle|^2 a_j) = \sum_i f(\langle f_i|A|f_i\rangle)$. $\square$ Now, that we defined a function of an operator, we are prepared to define the Von Neumann entropy. It is given by:

Definition 5.15 (Von Neumann Entropy)**.** *If a state is described by a trace one positive hermitian density matrix ρ, then its entropy is defined as $S(\rho) = -Tr(\rho\log\rho)$.*

If ρ is diagonal, then $S(\rho) = \sum_{i=1}^{n} \lambda_i \log \lambda_i$, *i.e.*, if the density matrix is diagonal, then the Von Neumann entropy is equal to the Shannon entropy of the vector formed by its eigenvalues. This is a positive quantity ($S(\rho) \geq 0$), as we will see later, and it is equal to 0 *iff* the state is a pure state, (see Theorem 5.18).

Wilde, in [67], says: "Why should the entropy of a pure quantum state vanish? It seems that there is quantum uncertainty inherent in the state itself and that a measure of quantum uncertainty should capture this fact. This last observation only makes sense if we do not know anything about the state that is prepared." Of course that if we prepared a pure state, we know exactly which measurements we need to perform in order to always obtain a certain result.

Let us suppose now that we have prepared the following pure state $|\psi\rangle\langle\psi|$. We know that we can always prepare an experimental device to detect this pure state by doing the following measurements: $\{|\psi\rangle\langle\psi|,\ \mathbb{I} - |\psi\rangle\langle\psi|\}$. As we know exactly how this state was prepared, we should not expect to learn anything from this measurement, because its outcome was known in advance. Therefore, of course, we can always perform a quantum measurement in order to verify which quantum state was prepared [67].

A statistical quantum operator can be seen as a description of a mixture of pure states. These pure states are given statistical operators (with rank one) and are represented by rays of a Hilbert space [53]. In 1927 Von Neumann associated an entropy function to a quantum operator and this discussion was also later extended in his book [53]. Let us assume that the density ρ is the mixture of orthogonal densities ρ_1 and ρ_2. In some textbooks, see for example [68], we can find Eqs. 5.5, 5.6 and 5.7 meaning the entropy of a two gases mixture. The total entropy of this mixture is equal to the sum of the two entropies and a well-defined entropy of the mixture [53].

$$\rho = p\rho_1 + (1-p)\rho_2, \tag{5.5}$$

$$pS(\rho_1) + (1-p)S(\rho_2) = S(\rho) - p\log p - (1-p)\log(1-p), \tag{5.6}$$

$$pS(\rho_1) + (1-p)S(\rho_2) = S(\rho) + H(p,\ (1-p)). \tag{5.7}$$

From this two component mixture we can generalize:

$$\rho = \sum_i \rho_i = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|, \tag{5.8}$$

$$S(\rho) = \sum_i \lambda_i S(|\psi_i\rangle\langle\psi_i|) - \sum_i \lambda_i \log \lambda_i. \tag{5.9}$$

Where the λ_i, ($0 \leq \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$), are the parameters of this convex mixture that they can always be interpreted as probabilities. Eq. 5.9 is the so-called *Schatten decomposition* of the mixed quantum state given by Eq. 5.8, and it reduces the determination of the (thermodynamic) entropy of a mixed state to the determination of the entropy of pure states [53, 69]. This decomposition is not unique, as we will see in the Theorem 5.32.

The argument presented by Von Neumann does not require the operator ρ to be a mixture of pure states [69]. We just need in all those equations above is the following property $\rho = p\rho_1 + (1-p)\rho_2$, in such a way that the

mixed states given by ρ_1 and ρ_2 are disjoint in the thermodynamic sense, in a two component gas mixture, for example, when there exists a completely permeable wall for all the molecules of a ρ_1-gas and isolated for the molecules of a ρ_2-gas [69]. Particularly, if these states ρ_1 and ρ_2 are disjoint, then this fact should be evidenced by a certain type of filter. The filter, that evidences the disjointedness of ρ_1 and ρ_2, can be expressed mathematically by the orthogonality of the eigenvectors of the two density matrices [53, 69].

Theorem 5.16. *Let ρ_1 and ρ_2 be two density matrices and $0 \leq p \leq 1$. The following inequality holds: $pS(\rho_1) + (1-p)S(\rho_2) \leq S(p\rho_1 + (1-p)\rho_2)$.*

Proof: This is true since $pS(\rho_1) + (1-p)S(\rho_2) = S(p\rho_1 + (1-p)\rho_2) + H(p, (1-p))$, and H is a positive quantity. □ Another interesting way to define the Von Neumann entropy is to define first the Shannon entropy of a density matrix relative to any POVM $\{E_i\}_i$: $H(\rho) \equiv H(\vec{p})$, where $p_i = Tr(E_i\rho)$. To make this definition independent of the chosen POVM, we need to minimize over all POVM's:

Definition 5.17 (Von Neumann Entropy)**.** *The entropy of Von Neumann can be defined as: $S(\rho) = min_{E_i \in POVM}\{H(\vec{p}),\ \ s.t.\ p_i = Tr(E_i\rho)\}$.*

This definition has an important special case, (Theorem 5.3).

5.5 Some Properties of the Von Neumann Entropy

Theorem 5.18 (Positivity of $S(\rho)$)**.** *The Von Neumann Entropy $S(\rho) \geq 0$, and the equality holds iff ρ is a pure state.*

Proof: Let ρ be a density matrix. It is a hermitian positive and trace one operator. Then it can be always be diagonalized. The positivity of $S(\rho)$ follows immediately from the positivity of the Shannon entropy, since the Von Neumann is defined as the Shannon entropy of the eigenvalues of ρ and $\rho \geq 0$. If ρ is a pure state, it has only one eigenvalue equal to 1 and all the other are equal to 0. As we defined $0 \log 0 \to 0$, by continuity reasons, then $S(\rho) = 0$ *iff* ρ is a pure state. □

Theorem 5.19 (Global Unitary Invariance)**.** $S(U\rho U^\dagger) = S(\rho)$.

Proof: It follows by observing that $U\rho U^\dagger = U \sum_i \lambda_i |e_i\rangle\langle e_i| U^\dagger = \sum_i \lambda_i |u_i\rangle\langle u_i|$, where $U|e_i\rangle = |u_i\rangle$. This property follows because the entropy is a function of the eigenvalues of ρ. □

5.5.1 Sub-additivity Theorems

Theorem 5.20 (Additivity of $S(\rho)$)**.** *Let us suppose we have $\rho_1 \in \mathcal{H}_1$ and $\rho_2 \in \mathcal{H}_2 \cdots \rho_n \in \mathcal{H}_n$. Then the entropy of $\rho = \rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n \in \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_n$ is $S(\rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n) = S(\rho_1) + \cdots + S(\rho_n)$, in case of n copies of ρ, $S(\rho^{\otimes n}) = nS(\rho)$.*

Proof: The proof is simple. Let $|i\rangle$ be the eigenvector of the i-th matrix with eigenvalue λ_i with $i = 1, \cdots, n$. Then $|1\rangle \otimes \cdots \otimes |n\rangle$, is the eigenvector

of $\rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n$, with eigenvalue $\prod_{i=1}^{n} \lambda_i$. Then $S(\rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n) = -\sum_i \prod_{i=1}^{n} \lambda_i \log(\prod_{i=1}^{n} \lambda_i) = -\sum_i (\lambda_1 \cdots \lambda_i \cdots \lambda_n) \sum_{i=1}^{n} \log(\lambda_i)$. Hence we have the following result: $S(\rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n) = S(\rho_1) + \cdots + S(\rho_2)$. □ Of course, if we have a product[7] of n density matrices ρ, then $S(\rho \otimes^1 \rho \otimes^2 \cdots \otimes^{n-1} \rho) = S(\rho^{\otimes n}) = nS(\rho)$. This theorem states that the information of the total system described by $\rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_n = \rho^{\otimes n}$, is equal to the sum of the information of its constituents, given that the total system can be written in a tensor product [66]. See Subsection 3.5.1 for more discussions about some properties of the tensor product and for other discussions over the *tensor product assumption*.

One relevant class of quantum inequalities identifies subsystems as a compound system, whose Hilbert space is given by a tensor product of the Hilbert spaces for the subsystems ($\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2$) [70]. Let us suppose that the state of a composite system is given by the following density matrix ρ_{12}. Therefore, the states of subsystems 1 and 2 will be given by their reduced density matrices, *e.g.*, $\rho_1 = Tr_2(\rho_{12}) = \mathbb{I}_1 \otimes Tr_2(\rho_{12})$, which can be obtained by taking the partial trace of ρ_{12} [70].

Theorem 5.21 (Sub-additivity of $S(\rho)$)**.** $S(\rho_{12}) \leq S(\rho_1) + S(\rho_2) = S(\rho_1 \otimes \rho_2)$.

Proof: Using the Corollary 5.13 and defining $\rho_{12} = A$, and $B = \rho_1 \otimes \rho_2 \equiv (\rho_1 \otimes \mathbb{I}_2)(\mathbb{I}_1 \otimes \rho_2)$, we will have [70]:

$$Tr[A(\log A - \log B)] = Tr_{12}\{\rho_{12}[\log(\rho_{12}) - \log(\rho_1 \otimes \rho_2)]\} =$$
$$= Tr_{12}\{\rho_{12}[\log(\rho_{12})]\} - Tr_{12}\{\rho_{12}[\log(\rho_1 \otimes \mathbb{I}_2)]\} - Tr_{12}\{\rho_{12}[\log(\mathbb{I}_1 \otimes \rho_2)]\} =$$
$$= Tr_{12}[\rho_{12}\log(\rho_{12})] - Tr_1[\rho_1 \log(\rho_1)] - Tr_2[\rho_2 \log(\rho_2)] =$$
$$= -S(\rho_{12}) + S(\rho_1) + S(\rho_2) \geq Tr(A - B) = Tr(\rho_{12} - \rho_1 \otimes \rho_2) \overset{a}{=} 0,$$
$$\text{Finally: } S(\rho_{12}) \leq S(\rho_1) + S(\rho_2) = S(\rho_1 \otimes \rho_2). \square$$
$$\text{The equality holds } \textit{iff } A = B, \text{ or: } \rho_{12} = \rho_1 \otimes \rho_2 = (\rho_1 \otimes \mathbb{I}_2)(\mathbb{I}_1 \otimes \rho_2). \square$$

Where in a, we use the property that $Tr(A \otimes B) = Tr(A)Tr(B)$, see for example [13].

Theorem 5.22 (Strong Sub-additivity of $S(\rho)$)**.** *If the composite state ρ_{123} is normalized, then $S(\rho_{123}) \leq S(\rho_{12}) + S(\rho_{23})$.*

Proof: Using the Corollary 5.13 and substituting $\rho_{123} = A$ and $B = e^{\log \rho_{12} + \log \rho_{23}}$, we will have [70]:

$$Tr[A(\log A - \log B)] = Tr(\rho_{123}\log\rho_{123}) - Tr[\rho_{123}(\log\rho_{12})] - Tr[\rho_{123}(\log\rho_{23})] =$$
$$= S(\rho_{123}) - S(\rho_{12}) - S(\rho_{23}) \geq Tr(A - B) = Tr\left[\rho_{123} - e^{(\log\rho_{12} + \log\rho_{23})}\right],$$
$$S(\rho_{123}) - S(\rho_{12}) - S(\rho_{23}) \geq 1 - Tr_{123}\left[e^{(\log\rho_{12} + \log\rho_{23})}\right],$$
$$S(\rho_{123}) - S(\rho_{12}) - S(\rho_{23}) \geq 1 - Tr_2(\rho_2)^2,$$
$$S(\rho_{123}) - S(\rho_{12}) - S(\rho_{23}) \overset{a}{\geq} 1 - Tr_2(\rho_2) = 0. \square$$

[7] The reader should not be tempted to think that this number n is large. In this text $\rho^{\otimes n} \approx \rho^{\otimes 8}$. This is a good notation for 1, 2, $\cdots$, 8 qubits. The reason why this number cannot be large is due to some obvious experimental reasons.

Where in a, we use the fact that $Tr(\rho)^2 \leq Tr(\rho)$, for any ρ [70]. The equality holds *iff* $A = B$, then $\rho_{123} = e^{\log \rho_{12} + \log \rho_{23}}$, or if $\rho_{123} = \rho_{12} \otimes \rho_{23}$. □

Theorem 5.23 (Concavity of $S(\rho)$)**.** *The entropy satisfies the following inequality:* $S(\sum_i p_i \rho_i) \geq \sum_i p_i S(\rho_i)$*, with* $0 \leq p_i \leq 1$ *and* $\sum_i p_i = 1$*. This inequality implies that Von Neumann entropy needs to be a concave function.*

Proof: (Due to [10]). The intuition is that the $\sum_i p_i \rho_i$ expresses the state of a quantum system which is in the state ρ_i with probability p_i and the uncertainty of the mixture of states should be higher than the average of the states ρ_i, since the state $\sum_i p_i \rho_i$ represents our ignorance of the quantum state [10]. Let us define the state $\rho_{AB} = \sum_i \rho_i \otimes |i\rangle\langle i|$. Then, we have $S(\rho_A) = S(\sum_i p_i \rho_i)$, and $S(\rho_B) = S(\sum_i p_i |i\rangle\langle i|) = H(p_i)$. Thus, $S(\rho_{AB}) = H(p_i) + \sum_i p_i S(\rho_i)$. Applying the Theorem 5.21 *i.e.*, $S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$, we have $\sum_i p_i S(\rho_i) \leq S(\sum_i p_i \rho_i)$ [10]. The concavity of the entropy follows immediately from the Jensen inequality (Theorem 4.17). □

Theorem 5.24 (Concavity of $S(\rho)$)**.** *If* $\rho = p\sigma + (1-p)\omega$*, then* $S(\rho) = S(p\sigma + (1-p)\omega) \geq pS(\sigma) + (1-p)S(\omega)$.

Proof: If $\rho = p\sigma + (1-p)\omega$, then by using Corollary 5.13 twice with $A = \sigma,\ \omega$ and $B = \rho$, we will have:

$$
\begin{aligned}
0 = Tr(A-B) \leq Tr[A(\log A - \log B)], \\
Tr[A(\log B)] \leq Tr[A(\log A)], \\
pTr[\sigma(\log \rho)] \overset{a}{\leq} pTr[\sigma(\log \sigma)], \\
(1-p)Tr[\omega(\log \rho)] \overset{b}{\leq} (1-p)Tr[\omega(\log \omega)], \\
-S(\rho) = Tr(\rho \log \rho) = pTr(\sigma \log \rho) + (1-p)Tr(\omega \log \rho) \overset{c}{\leq} \\
\overset{c}{\leq} pTr(\sigma \log \sigma) + (1-p)Tr(\omega \log \omega) = -pS(\sigma) - (1-p)S(\omega), \\
S(\rho) \overset{d}{\geq} pS(\sigma) + (1-p)S(\omega). \ \square
\end{aligned}
$$

The left-hand of the inequality c, is the sum of the left-hand of the inequalities a and b. Analogously, we have the same for the right-hand of the inequality c. A sign reversion (d) gives us the result. Then it is easy to understand that Von Neumann entropy is a concave function: $S(E(\rho)) \geq E(S(\rho))$.

But why is concavity considered to be important? We already know that entropy is a measure of lack of information. Hence, if two densities are put together in an ensemble (in a convex combination $\sum_i p_i \rho_i$), one loses the information that tells from which density a special sample comes from, and therefore entropy increases [66].

5.5.2 Other Properties of the Von Neumann Entropy

Proposition 5.25. *There exists a set of unitary matrices* $\{U_k\}$ *and a vector of probabilities* $\vec{p}$*, such that for any hermitian matrix* ρ*, with* $Tr(\rho) = 1$ *we always have* $\sum_k p_k U_k \rho U_k^\dagger = \frac{\mathbb{I}}{n}$.

Proof: This proof is straightforward.

The Maximal Value for the Von Neumann Entropy

The quantum determinism postulate asserts that a quantum system prepared in a pure state remains pure when it evolves in a perfectly controlled environment [15]. Suppose that we have prepared a pure quantum state ρ, then the Von Neumann entropy of ρ must be invariant under unitary evolutions, and this is true since $S(\rho) = -Tr(\rho \log \rho)$ remains invariant under unitary evolution in a perfectly known isolated system, because $\tilde{\rho} = U\rho U^\dagger$ in such evolutions, and $S(\tilde{\rho}) = -Tr[U\rho U^\dagger \log(U\rho U^\dagger)] = S(\rho)$, because of the cyclic property of the trace function.

If the environment is partially known, we must replace the global unitary U by an ensemble of unitary matrices U_k with respective probabilities p_k. Then the evolution can be described by $\tilde{\rho} = \sum_k p_k U_k \rho U_k^\dagger$. Then $S(\tilde{\rho}) = -Tr(\tilde{\rho} \log \tilde{\rho}) \geq S(\rho) = -Tr(\rho \log \rho)$ [15]. This result will be discussed properly in the following theorem, (Theorem 5.26).

Theorem 5.26 (Maximal value for $S(\rho)$)**.** *$S(\rho) \leq \log n$. For any state ρ.*

Proof: Since we know by the Prop. 5.25 that, for any density matrix ρ, there exists a vector of probabilities and a set of unitary matrices $\{U\}$ such that: $\sum_k p_k U_k \rho U_k^\dagger = \frac{\mathbb{I}}{n}$, and by using Theorem 5.23, (the Jensen inequality for concave functions, see Theorem 4.17), we have that $\sum_k p_k S(U_k \rho U_k^\dagger) \overset{a}{=} S(\rho) \leq S(\sum_k p_k U_k \rho_k U_k^\dagger) = S(\frac{\mathbb{I}}{n}) = \log n$. Because globally unitary evolutions do not change the entropy, due to the cyclic property of the trace function (equality *a*). □

Another Proof: Another proof can be obtained if we remember that the quantum relative function D is always a positive function, that is $D \geq 0$, (see Section 5.9). Therefore, $D(\rho || \frac{\mathbb{I}}{n}) \geq 0$. This implies that $D(\rho || \frac{\mathbb{I}}{n}) = \log n - S(\rho) \geq 0$, hence $S(\rho) \leq \log n$. □

Theorem 5.27. *Let ρ and σ be two density matrices, and let us suppose that $\rho \prec \sigma$. Then $S(\rho) \leq S(\sigma)$.*

Proof: By using Uhlmann's theorem, (Theorem 5.1), there exists a set of unitary matrices $\{U\}$ and a probability vector $\vec{p}$, such that $\rho = \sum_i p_i U_i \sigma U_i^\dagger$. By using the concavity of S, we have $S(\rho) \leq \sum_i p_i S(U_i \sigma U_i^\dagger) = S(\sigma)$. □

Thus, if a probability distribution or a density matrix is more disordered than another in the sense of majorization, then they are also according to the Shannon and Von Neumann entropies [38]. We already know that any function that also preserves the majorization are called Schur convex (or concave), see Def. 4.11.

5.6 Some Interpretations of the Expression for the Entropy

In 1877, Boltzmann established the connection between the variable of state *entropy* and the disorder of a physical system in his famous formula $S = k \log W$, where W means the number[8] of micro-states which have the same

[8] In German the W means the *thermodynamische Wahrscheinlichkeit* – thermodynamic probability.

macroscopic properties [66]. This quantity requires a better interpretation in classical mechanics, but, in quantum mechanics, this number is a well-defined quantity: the number of micro-states may be interpreted as the number of pure states with some prescribed expectation values [66]. This lack of interpretation for the entropy in classical mechanics does not occur in quantum theory because of its probabilistic nature.

Let us suppose now that we have a rank r density matrix ρ written in its eigenstates $\{|e_i\}\rangle$. Then $\rho = \sum_{i=1}^{r} \lambda_i |e_i\rangle\langle e_i|$, where λ_i, is the probability to find the state in the pure state $\{|e_i\rangle\}$, with $\sum_i \lambda_i = 1$, and $0 \leq \lambda_i \leq 1$, $\forall i$. If we perform n measurements (n large) of identical copies of ρ, the system will be found in the state $|e_1\rangle$, $\lambda_1 n$ times, in the state $|e_2\rangle$, $\lambda_2 n$ times so forth, and of course this numbers must be integers [66]. The matrix ρ does not contain information about the ordering of the states $\{|e_1\rangle, |e_2\rangle, \cdots, |e_r\rangle\}$. There exists $W_n = \frac{n!}{(\lambda_1 n)!(\lambda_2 n)!\cdots(\lambda_r n)!}$ possibilities for this. But for n large, we know that by using the Stirling formula, $\frac{1}{n}\log(\frac{n!}{(\lambda_1 n)!(\lambda_2 n)!\cdots(\lambda_r n)!})$ converges to S [66].

This number of micro-states, in quantum information theory, can be interpreted in the following manner: let us consider n copies of ρ, representing a total Hilbert space $\mathcal{H}_{\mathcal{HS}} \otimes \mathcal{H}_{\mathcal{HS}} \otimes \cdots \otimes \mathcal{H}_{\mathcal{HS}}$, where $\mathcal{H}_{\mathcal{HS}}$ represents the space for the original system [66]. In this new system, there are micro-states of the form $|e_1\rangle \otimes |e_2\rangle \otimes \cdots \otimes |e_r\rangle$, and all the micro-states have the same weight, because $|e_1\rangle$ occurs $\lambda_1 n$ times, and $|e_2\rangle$ occurs $\lambda_2 n$ times so forth [66]. The entropy is $\log W_n$, with $W_n = \frac{n!}{(\lambda_1 n)!(\lambda_2 n)!\cdots(\lambda_r n)!}$, then $\frac{1}{n}\log W_n \to S$, when $n \to \infty$, see for example [66, 69]. This limit can be calculated easily, remembering that $\lim_{n\to\infty} \frac{1}{n}(\log W_n) = \lim_{n\to\infty}(\frac{1}{n}\log \frac{n!}{(\lambda_1 n)!(\lambda_2 n)!\cdots(\lambda_r n)!})$, by using Stirling's approximation $\log n! \approx n\log n - n$, and keeping the most important terms, we have $\frac{1}{n}[n\log n - \sum_{i=1}^{n}(\lambda_i n)\log(\lambda_i n)] = \frac{1}{n}[\sum_{i=1}^{n}(\lambda_i n)\log n - \sum_{i=1}^{n}(\lambda_i n)\log(\lambda_i n)] = -\sum_{i=1}^{n}\lambda_i \log\lambda_i = S(\rho)$.

In information theory, entropy is a measure of our ignorance towards a system described as a density matrix. The formal correspondence between the Shannon entropy and the $S(\rho) = -\sum_i \lambda_i \log\lambda_i$ is obvious, but we need to derive some theorems first to understand it. Later this correspondence will be fully understood with the help of the mixture theorem (Theorem 5.32).

If we try to adapt the classical concepts to the quantum ones, we will find that the perfect analogy does not exist. Accordingly to the usual *dictionary*, we would have to replace [66]:

- A subset of the phase space *by* a projection.
- A measure of a set *by* a trace (measure of the corresponding projection).
- A density distribution (as discussed in [25]) *by* a density matrix.

5.7 States with Maximum Von Neumann Entropy

If the energy of the system is fixed, we know that the state in the micro-canonical ensemble is the equilibrium state. This is always argued on philosophical grounds [28, 52, 66], and it is a direct application of the Laplace's

principle of insufficient reason for physical systems. This is only possible because the equilibrium states enjoy some remarkable stability properties which can be characterized as follows: small perturbations in equilibrium systems can change the state of the system only locally, but not globally. This, of course, cannot be obtained in a time invariant, but not within an equilibrium system, because an arbitrarily small perturbation may be sufficient to produce a transition to a totally different state [66].

Let us state the following problem: Given a fixed Hamiltonian H with $\langle E \rangle = Tr(\rho H)$. What is the density matrix ρ with the maximum value of $S(\rho)$? The answer is well-known, and it is given by Theorem 5.29.

Definition 5.28 (The Gibbs Quantum State). $\sigma_G = \frac{e^{-\beta H}}{Z}$, *with* $Z = Tr(e^{-\beta H})$. *Where* β *is chosen such that* $Tr(\sigma_G(\beta)H) = \langle E \rangle$.

Theorem 5.29. *If* σ_G *is the Gibbs quantum state, then* $S(\sigma_G) \geq S(\rho)$ *for any* ρ *that satisfies the constraint* $\langle E \rangle = Tr(\rho H) = Tr(\sigma_G H)$.

Proof: We are interested in computing $sup_{\{\rho \in M_n(\mathcal{C})\}}\{S(\rho),\ \text{s.t. } Tr(\rho H) = \langle E \rangle \equiv E,\ \text{with } \rho \geq 0,\ Tr(\rho) = 1\}$ [71]. Let us suppose that $Tr(\rho H) \leq E$ and $Tr(\sigma_G H) = E$ (the inequality a is due this assumption). Then, by using the inequality given by the Corollary 5.13 (in inequality b), we obtain:

$$
\begin{aligned}
Tr[\rho(\log \sigma_G)] &= -\beta Tr(\rho H) - \log[Tr(e^{-\beta H})], \\
Tr[\sigma_G(\log \sigma_G)] &= -\beta Tr(\sigma_G H) - \log[Tr(e^{-\beta H})], \\
-Tr[\rho(\log \sigma_G)] &\overset{a}{\leq} -Tr[\sigma_G(\log \sigma_G)], \\
S(\rho) = -Tr[\rho(\log \rho)] &\overset{b}{\leq} -Tr[\rho(\log \sigma_G)] \overset{a}{\leq} -Tr[\sigma_G(\log \sigma_G)] = S(\sigma_G), \\
S(\sigma_G) &= sup_{\{\rho \in M_n(\mathcal{C})\}}\{S(\rho),\ \text{s.t. } Tr(\rho H) = E,\ \text{with } \rho \geq 0,\ Tr(\rho) = 1\}. \ \square
\end{aligned}
$$

5.8 The Set of Mixed States

We have already shown in Chapter 3 that the space of the n^2 dimensional density matrices $\mathcal{M}^n$ is a convex set. This set is the intersection of the space of the hermitian matrices with a hyperplane parallel to the linear subspace of traceless operators. The pure states are projections onto one-dimensional subspaces in the Hilbert space $\mathcal{H}$ [31].

We can define two different ways to write some mixed states. The Eq. 5.30 is an exponential representation of a density matrix. The next equation represents a $SU(n)$ expansion. For one qubit, we have already showed in Chapter 2, that the Pauli matrices form a basis and they are members of $SU(2)$, then the expansion presented in Eq. 5.31 looks like a generalized Bloch representation. Hence, we will define two formulas for writing expansions for a quantum state:

Definition 5.30 (Exponential Coordinates). $\rho = \frac{e^{-\beta H}}{Tr(e^{-\beta H})}$.

Definition 5.31 ($SU(n)$ basis). $\rho = \frac{\mathbb{I}}{n} + \sum_{i=1}^{n^2-1} r_i \sigma_i$.

As we showed in the Chapter 2, the maximally mixed state $\mathbb{I}_* = \frac{\mathbb{I}}{n}$ occurs when the *Bloch's* vector $|\vec{r}| = 0$ (Def. 5.31). For one qubit, this matrix, which is the trace one identity matrix, also corresponds to the unpolarized radiation.

A convex combination of density matrices lies inside the line formed by these matrices. Then, by using Def. 3.27 and Def. 5.31, we can show that the Hilbert-Schmidt distance is a kind of an *euclidean* distance[9] [31]:

$$D_2(\rho, \tilde{\rho}) = \sqrt{\frac{1}{2} Tr\{\sum_{i,j}[(r_i - \tilde{r}_j)\sigma_i\sigma_j]^2]\}},$$
$$D_2(\rho, \tilde{\rho}) = \sqrt{\sum_i (r_i - \tilde{r}_i)^2}.$$

5.8.1 The Schrödinger Mixture Theorem - 1936

Let us consider a measurable physical quantity represented by a hermitian operator $\mathcal{O}$. If the system is prepared in the following state $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, we have already seen that the expectation value of the observable $\mathcal{O}$, is given by $\langle\mathcal{O}\rangle = Tr(\mathcal{O}\rho)$. The probability p_o, that a measurement of $\mathcal{O}$ yields a particular eigenvalue λ_o, is also expressible as an expectation value. If we define first the projector Π_o, which projects into the subspace related to the chosen eigenvalue, then we can say that $p_o = Tr(\rho\Pi_o)$.

Now, suppose that a quantum state is prepared in a pure state given by $|\psi_i\rangle$. Then the expected value of this physical quantity is given by $\langle\mathcal{O}\rangle_i = \langle\psi_i|\mathcal{O}|\psi_i\rangle$. Averaging over the probability distribution, we have $\langle\mathcal{O}\rangle = \sum_i p_i \langle\mathcal{O}\rangle_i = Tr(\mathcal{O}\rho)$.

Let us make a break for a moment with the usage of the Dirac notation in order to motivate the Schrödinger mixture theorem. A quantum system ψ_i can be written in terms of a complete orthonormal set of functions u_k. Then $\psi_i = \sum_k a_{ki} u_k$, and the normalization condition implies that $\sum_k |a_{ki}|^2 = 1$ [60]. Let us come back to the expectation value calculated before: $\langle\mathcal{O}\rangle_i = \sum_{kn} a_{ki} a^*_{ni} \mathcal{O}_{nk}$, where $\mathcal{O}_{nk} = \langle u_n, \mathcal{O} u_k\rangle$ are the matrix elements in the basis u_k and the density matrix is $\rho_{kn} = \sum_i p_i a_{ki} a^*_{ni}$. This matrix can also be interpreted as an expected value over the probability distribution $\{p_i\}$, then $\rho_{kn} = E(a_{ki} a^*_{ni})_{\{p_i\}}$ [60].

If we observe this expectation value $\rho_{kn} = E(a_{ki} a^*_{ni})_{\{p_i\}}$, we can suppose that an infinite number of different arrays exists, representing different mixtures of pure states, all leading to the same density matrix [60]. The most general discrete array which leads to a given density matrix ρ corresponds to:

$$\rho = AA^\dagger. \tag{5.10}$$

Where A is a possible non-square matrix. An array is defined by its matrix elements $A_{ki} = \sqrt{p_i} a_{ki}$, with $\sum_k |A_{ki}|^2 = p_i$. To find another array that corresponds to the same density matrix, just insert a unitary matrix U [60]:

$$\rho = (AU)(U^{-1}A^\dagger). \tag{5.11}$$

[9] Here we used the following property: $Tr(\sigma_i\sigma_j) = 2\delta_{ij}$ [31].

The Eq. 5.11 has the form $\rho = BB^\dagger$, with $B = AU$, since U is a unitary matrix [60]. Supposing that we have a set of n unitary matrices $\{U_i\}_{i=1}^n$, then if the following matrix products $U_1U_1^\dagger = \cdots = U_nU_n^\dagger = \mathbb{I}$ are representations of the identity matrix, it is easy to perceive that we can represent a density matrix as $\rho = AA^\dagger = A\mathbb{I}A^\dagger = AU_1U_1^\dagger A^\dagger = \cdots = AU_nU_n^\dagger A^\dagger$, because the identity matrix can always be written as a product of *any* unitary matrix U_i by its inverse, *i.e.*, its complex conjugate matrix ($U_i^\dagger$). Then the number of products can be infinite, thus a density matrix can be written in an infinite number of ways. Note that the internal dimension k of the matrix products $\rho_{n\times n} = (A_i)_{nk}(A_i^\dagger)_{kn}$ can also change, but the external dimension n cannot, in order to preserve the $n \times n$ dimension of the density matrix $\rho_{n\times n} = \sum_k A_{nk}A_{kn}^\dagger$. This result is known as the Schrödinger mixture theorem (Theorem 5.32), and it is proved now with the help of the Dirac notation.

Theorem 5.32 (The Schrödinger Mixture Theorem)**.** *A density matrix having the diagonal form $\rho = \sum_{i=1}^r \lambda_i|e_i\rangle\langle e_i|$ can be written in a statistical mixture of M operators,* i.e., *$\rho = \sum_{i=1}^M p_i|\psi_i\rangle\langle\psi_i|$, iff there exists a unitary matrix U, ($dim(U) = M^2$), such that: $|\psi_i\rangle = \frac{1}{\sqrt{p_i}}\sum_j^r U_{ij}\sqrt{\lambda_i}|e_i\rangle$* [31].

Proof: (Due to [31]) First observe that the matrix U does not act on the Hilbert-Schmidt space, because we can have $M > r$. But only the first r columns are needed here. The remaining $M - r$ columns are just added in order to build a unitary matrix. To prove the theorem, just define the first r columns of U as:

$$U_{ij} \equiv \sqrt{\frac{p_i}{\lambda_j}}\langle e_j|\psi_i\rangle \Rightarrow \sum_{j=1}^r U_{ij}\sqrt{\lambda_j}|e_j\rangle = \sqrt{p_i}|\psi_i\rangle. \tag{5.12}$$

Proposition 5.33. *The matrix U defined in Eq. 5.12 is unitary.*

Proof: A unitary matrix is a complex matrix U satisfying the condition: $UU^\dagger = U^\dagger U = \mathbb{I}$. Note that this condition implies that a matrix U is unitary *iff* it has an inverse which is equal to its conjugate transpose, *i.e.*, $U^{-1}U = \mathbb{I} = U^\dagger U$, then $U^{-1} = U^\dagger$. Let $\mathbb{I}$ be the identity matrix defined by its matrix elements $(\mathbb{I})_{ij} = \delta_{ij}$. Then, using the definition of the matrix elements of U, $U_{ij} \equiv \sqrt{\frac{p_i}{\lambda_j}}\langle e_j|\psi_i\rangle$ and using the fact that its inverse is equal to $U^\dagger$, we get the proof:

$$\begin{aligned}
(\mathbb{I})_{ij} = (U^\dagger U)_{ij} &= \sum_{k=1}^M U_{ik}^\dagger U_{kj}, \\
&= (U^\dagger U)_{ij} = \sum_{k=1}^M (U^*)_{ki}U_{kj}, \\
&\overset{a}{=} \sum_{k=1}^M \frac{p_k}{\sqrt{\lambda_i\lambda_j}}\langle\psi_k|e_i\rangle\langle e_j|\psi_k\rangle, \\
&= \frac{1}{\sqrt{\lambda_i\lambda_j}}\langle e_j|\psi_k\rangle\sum_{k=1}^M p_k\langle\psi_k|e_i\rangle,
\end{aligned}$$

$$\stackrel{b}{=} \frac{1}{\sqrt{\lambda_i\lambda_j}}\langle e_j|\rho|e_i\rangle,$$
$$\stackrel{c}{=} \frac{1}{\sqrt{\lambda_i\lambda_j}}\langle e_j|(\sum_{i=1}^{r}\lambda_i|e_i\rangle\langle e_i|)|e_i\rangle,$$
$$\stackrel{d}{=} \frac{\lambda_i}{\sqrt{\lambda_i\lambda_j}}\langle e_j|e_i\rangle,$$
$$(\mathbb{I})_{ij} = \delta_{ij}. \square$$

In a, we use the definition of U, in b, we use the expansion of ρ in the basis $\{|\psi_i\rangle\}$, and in c, d we use the fact that the eigenstates $|e_i\rangle$ are normalized, that is $\langle e_i|e_i\rangle = 1$, and eigenvectors related to different eigenvalues in a hermitian matrix are always orthogonal, *i.e.*, $\langle e_j|e_i\rangle = \delta_{ij}$. Now we are prepared to demonstrate the Theorem 5.32:

$$\rho = \sum_{i=1}^{M} p_i|\psi_i\rangle\langle\psi_i|,$$
$$= \sum_{i=1}^{M}\sqrt{p_i}|\psi_i\rangle\sqrt{p_i}\langle\psi_i|,$$
$$= \sum_{i=1}^{M} p_i\sum_{j=1}^{r}\frac{1}{\sqrt{p_i}}U_{ij}\sqrt{\lambda_j}|e_j\rangle\frac{1}{\sqrt{p_i}}U_{ji}^{*}\sqrt{\lambda_j}\langle e_j|,$$
$$= \sum_{j=1}^{r}(\sum_{i=1}^{M}U_{ij}U_{ij}^{*})\sqrt{\lambda_j\lambda_j}|e_j\rangle\langle e_j|,$$
$$\rho = \sum_{j=1}^{r}\lambda_j|e_j\rangle\langle e_j|. \square$$

Hence if $\rho = \sum_{i=1}^{r}\lambda_i|e_i\rangle\langle e_i|$, and if we define an unitary matrix such as $U_{ij} \equiv \sqrt{\frac{p_i}{\lambda_j}}\langle e_j|\lambda_i\rangle$ and $|\psi_i\rangle = \frac{1}{\sqrt{p_i}}\sum_j^r U_{ij}\sqrt{\lambda_i}|e_i\rangle$, we can write ρ as the following mixture: $\rho = \sum_{i=1}^{M} p_i|\psi_i\rangle\langle\psi_i|$. What have we proved here? The theorem 5.32 says that we can write a quantum state as a convex mixture of operators in an infinite number of ways, *i.e.*, we can define a density matrix ρ performing a preparation of an infinite number of operator ensembles, each one described by the operators weighted by a probability distribution $\{\vec{p} = \{p_i\}, |\psi_i\rangle\langle\psi_i|\}$.

We already know that the *entropy of a mixture* $H_{mix}(\vec{p})$ can be defined as the Shannon entropy of the probability distribution $\vec{p}$ of the mixture (see Section 4.2, Def. 4.13), where the $p_i = tr(|\psi_i\rangle\langle\psi_i|\rho)$ and the set $\{|\psi_i\rangle\langle\psi_i|\}_{i=1}^{M}$ are M operators that over-span the d^2-dimensional Hilbert-Schmidt space where the density matrix lives. Given all these operators $|\psi_i\rangle\langle\psi_i| \geq 0$, we can ensure that all probabilities $p_i \geq 0$. If ρ is written in its eigenstates, as discussed in Section 5.4, the Von Neumann entropy can be defined as the Shannon entropy of the matrix eigenvalues λ_i, *i.e.*, $S(\rho) = -tr(\rho\ln\rho) = -\sum_{i=1}^{r}\lambda_i\ln(\lambda_i)$. Then, we can state a Corollary for the Theorem 5.32:

Corollary 5.34. $H_{mix}(\vec{p}) \equiv -\sum_{i=1}^{M} p_i\log(p_i) \geq -\sum_{i=1}^{r}\lambda_i\log(\lambda_i) \equiv S(\rho)$, *and the equality holds iff* $M = r$, *i.e., if the mixture is the eigen-mixture.*

Proof: A quantum state ρ can be represented by a mixture of pure states[10] in many different ways. The most general representation of this density matrix ρ is $\rho = \sum_{i=1}^{M} p_i |\psi_i\rangle\langle\psi_i|$, as we have shown with the aid of Theorem 5.32. These various states $|\psi_i\rangle\langle\psi_i|$ might not be mutually orthogonal since we have only one way to expand a known density matrix with rank r with exactly r terms in the matrix eigenstates: $\rho = \sum_{i=1}^{r} \lambda_i |e_i\rangle\langle e_i|$, where the $\vec{\lambda} = \{\lambda_i\}$ are the eigenvalues of ρ, and the $\{|e_i\rangle\}$, its eigenvectors. First we need to show that this matrix (B) is bistochastic, and it relates the eigenvalue vector $\vec{\lambda}$ to the probability vector $\vec{p} = \{p_i\}$. We can see this fact if we multiply the equation 5.12 by $\sum_{l=1}^{r} \sqrt{\lambda_l}\langle e_l|U_{lk}^{\dagger} = \sqrt{p_k}\langle\psi_k|$:

$$p_i = \sum_{j=1}^{r} \left|U_{ij}\right|^2 \lambda_j, \tag{5.13}$$

$$\vec{p} = B\vec{\lambda}. \tag{5.14}$$

Then, by using Def. 4.3, we show that the matrix $B = (\left|U_{ij}\right|^2)$ is a bistochastic matrix and $\vec{p} = B\vec{\lambda}$. Using the Theorem HLP (Theorem 4.7), we can say that the vector $\vec{p}$ is majorized by the eigenvalue vector $\vec{\lambda}$. In other words, $\vec{p} \prec \vec{\lambda}$ [31]. We have already shown that $\vec{p} \prec \vec{\lambda}$. Both entropy functions H_{mix} and S are Schur concave functions (see Def. 4.11 and Theorem 4.45). If $\vec{x} \prec \vec{y}$ and $f(\star)$ is a Schur concave function, then $f(\vec{x}) \geq f(\vec{y})$. This proves the Corollary [31]. So, it is reasonable to say that $H_{mix} \geq S(\rho)$ and the equality holds only when the projectors are the eigenstates of ρ, *i.e.*, $M = r$. □ We do not need a rigorous proof for the Corollary 5.34 to understand that it is true. A good way to see this fact is to try to make a valid mixture of infinite terms of $|\psi_i\rangle\langle\psi_i|$, *i.e.*, to make $M \to \infty$, since $\sum_{i=1}^{\infty} p_i = 1$. This *artificial* mixture tells to us that $H_{mix} \to \infty$ is clearly bigger than $S(\rho)$, that is limited by construction.

Theorem 5.35 (Entropy of Preparation)**.** *If we prepare a state in the following ensemble: $\{|\psi_i\rangle,\ P = \{p_i\}\}_{i=1}^{n}$, so that the density matrix is given by $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, then $H(P) \geq S(\rho)$.*

Proof: By Theorem 5.32, the equality holds *iff* all $|\psi_i\rangle$ are mutually orthogonal. This Theorem states that the distinguishable is lost when we mix non-orthogonal states, then we cannot recover the information on which state it was prepared [50]. □

Theorem 5.36 (Entropy of Measurement)**.** *Suppose that the system is in the state ρ and we measure the observable $\mathcal{O} = \sum_i o_i |o_i\rangle\langle o_i|$. By the quantum mechanics postulates, the outcome o_i occurs with probability equal to $p(o_i) = Tr(\mathcal{O}\rho)$. Then the Shannon entropy of the ensemble of measurement outcomes $M_i = \{o_i,\ p(o_i)\}$ satisfies $H(M) \geq S(\rho)$* [50].

Proof: $H(M) = -\sum_i o_i \log o_i \geq -Tr(\rho \log \rho) = -\sum_i \lambda_i \log \lambda_i = S(\rho)$. This is true according to Theorem 5.32 and the equality holds *iff* $[\mathcal{O}, \rho] = 0$. □ Physically, it says that the randomness of the measurement results is always minimized if we elect to measure observables that commute with ρ [50].

[10] Pure states are rank one and trace one positive projectors ($P_i \equiv |\psi_i\rangle\langle\psi_i|$), such as $P_i^n = P_i$. See Definitions 2.8 and 3.22 and Theorem 3.24.

5.9 The Quantum Relative Entropy $D(P||Q)$

In Chapter 4, we have seen that the classical relative entropy $D(P||Q)$ played a key role as a measure of how different two probability distributions are from each other [31]. We need to define an analogue quantity for the quantum case: the quantum relative entropy $D(\rho||\sigma)$. It plays a similar role in the quantum information theory [31].

Definition 5.37 (Quantum Relative Entropy)**.** *For any pair of density matrices ρ and σ, the quantum relative entropy is defined as:* $D(\rho||\sigma) = Tr[\rho(\log\rho - \log\sigma)]$.

5.9.1 Some Properties of the Quantum Relative Entropy $D(P||Q)$

Theorem 5.38 (Positivity of D)**.** $D(\rho||\sigma) \geq 0$ *for any ρ and σ. The equality holds iff $\rho = \sigma$.*

Proof: By Corollary 5.13 with $A = \rho$ and $B = \sigma$. The equality holds *iff* $A = B \Rightarrow \rho = \sigma$, then $Tr(A\log A - A\log B) = Tr(\rho\log\rho - \rho\log\sigma) \equiv D(\rho||\sigma) \geq Tr(A - B) = Tr(\rho - \sigma) = 0$. We could prove this fact by using the following inequality $S(\rho||\sigma) \geq \frac{1}{2}Tr(\rho - \sigma)^2 = D_2^2(\rho, \sigma) \geq 0$. This inequality is an analogue to the classical inequality proved in Theorem 4.37. □ Let us suppose that we have one qubit and we want to calculate the quantum relative entropy between a pure state ρ and a mixed state defined as $\sigma = \varepsilon\rho + (1 - \varepsilon)\rho_\perp$. Then $D(\rho||\sigma) = Tr[\rho(\log\rho - \log\sigma)] = -\log\varepsilon$. Then $\lim_{\varepsilon\to 0} D(\rho||\sigma) = +\infty$.

Theorem 5.39 (Global Unitary Invariance)**.** $D(\rho_1||\rho_2) = D(U\rho_1U^\dagger||U\rho_2U^\dagger)$.

Proof: The proof is simple by using the cyclic property of the trace function. Then $D(U\rho_1U^\dagger||U\rho_2U^\dagger) = Tr[U\rho_1U^\dagger(\log U\rho_1U^\dagger - \log U\rho_2U^\dagger)] = D(\rho_1||\rho_2)$. □

Theorem 5.40 (Concavity of $D(\rho||\sigma)$)**.** $D(\rho||\sigma)$ *is a convex function.*

Proof: The proof is out of the scope of this text and the reader can find it in [71]. If ρ and σ commute, they can be diagonalized simultaneously by some unitary matrix U. This problem becomes identical to the classical relative entropy problem and it was proved in Theorem 4.36. □

Theorem 5.41 (Additivity Property)**.** $D(\rho_1\otimes\rho_2||\sigma_1\otimes\sigma_2) = D(\rho_1||\sigma_1) + D(\rho_2||\sigma_2)$.

Proof: Applying the above additivity relation inductively, it is possible to conclude that $D(\rho^{\otimes n}||\sigma^{\otimes n}) = nD(\rho||\sigma)$. □

Theorem 5.42 (Monotonicity Under Partial Trace)**.** $D[Tr_B(\rho_{AB})||Tr_B(\sigma_{AB})] \leq D(\rho_{AB}||\sigma_{AB})$.

Proof: (Due to [10], for another proof, see for example [70]). In Proposition 5.25, we showed that there exists unitary matrices given by U_j on the space $\mathcal{H}_B$ and a vector of probability $\vec{p}$ such that: $Tr_B(\rho_{AB}) = \rho_A \otimes \frac{\mathbb{I}}{n} = \sum_j p_j U_j\rho_{AB}U_j^\dagger$,

with $\rho_A \in \mathcal{H}_A$, $U \in \mathcal{H}_B$ and $\rho_A \otimes \frac{\mathbb{I}}{n} \in \mathcal{H}_{AB}$ for all ρ_{AB} [10]. Therefore, by using this proposition we will have:

$$D[(\rho_A \otimes \tfrac{\mathbb{I}}{n})||(\sigma_A \otimes \tfrac{\mathbb{I}}{n})] \overset{a}{\leq} \sum_j p_j D[(U_j \rho_{AB} U_j^\dagger)||(U_j \sigma_{AB} U_j^\dagger)],$$

$$D[(\rho_A \otimes \tfrac{\mathbb{I}}{n})||(\sigma_A \otimes \tfrac{\mathbb{I}}{n})] \overset{b}{\leq} \sum_j p_j D[(\rho_{AB})||(\sigma_{AB})] = D(\rho_{AB}||\sigma_{AB}),$$

$$D[Tr_B(\rho_{AB})||Tr_B(\sigma_{AB})] = D[(\rho_A \otimes \tfrac{\mathbb{I}}{n})||(\sigma_A \otimes \tfrac{\mathbb{I}}{n})] \leq D(\rho_{AB}||\sigma_{AB}). \square$$

The inequality a holds by the concavity of the quantum relative entropy (Theorem 5.40). The inequality b holds because the entropy is invariant under global unitaries, Theorem 5.39.

Theorem 5.43 (Monotonicity of Quantum Relative Entropy)**.** *The quantum relative entropy between two states ρ and σ can only decrease if we apply the same noisy map $\mathcal{E}$ to each state* [67].

Proof: A noise map acting in $(\star)$ is a map such that: $\mathcal{E}(\star) = Tr_E[U_{(\star)\otimes E}(\star) \otimes |E\rangle\langle E| U^\dagger_{(\star)\otimes E}]$, see subsection 5.3. Where $|E\rangle$ is the environment and $U_{(\star)\otimes E} = U_G$ is a global unitary [67]:

$$D(\rho||\sigma) \overset{a}{=} D(\rho||\sigma) + 0,$$

$$D(\rho||\sigma) \overset{b}{=} D(\rho||\sigma) + D[(|E\rangle\langle E|)||(|E\rangle\langle E|)],$$

$$D(\rho||\sigma) \overset{c}{=} D[(\rho \otimes |E\rangle\langle E|)||(\sigma \otimes |E\rangle\langle E|)],$$

$$D(\rho||\sigma) \overset{d}{=} D[U_G(\rho \otimes |E\rangle\langle E|)U_G^\dagger||U_G(\sigma \otimes |E\rangle\langle E|)U_G^\dagger],$$

$$D(\rho||\sigma) \overset{e}{\geq} D[\mathcal{E}(\rho)||\mathcal{E}(\sigma)]. \square$$

In the equality a, we just sum 0. In equality b, we write $0 = D[(|E\rangle\langle E|)||(|E\rangle\langle E|)]$. The equality c follows from additivity of quantum relative entropy over the tensor product states [67]. The equality d follows because D is invariant under global unitaries (Theorem 5.39). The inequality e is the monotonicity under partial trace, (Theorem 5.42). It follows easily from the given simpler form of monotonicity: $D(\rho_{AB}||\sigma_{AB}) \geq D(\rho_A||\sigma_A)$. This last inequality is intuitive, since when we trace a part of the system, it becomes less distinguishable.

Theorem 5.44 (The Convergence in Quantum Relative entropy)**.** *The convergence in quantum relative entropy implies convergence in the trace norm $\|\star\|_1$.*

For one qubit states ρ and σ that are diagonal in the same basis, *i.e.*, $\rho = p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|$ and $\sigma = q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|$, the proof is identical to those proofs in the Subsection 4.4.3, (Lemma 4.40 and Theorem 4.41). For a more interesting version of this theorem, let us consider the projector P onto the positive eigenstate of $\rho - \sigma$, and let $\mathbb{I} - P$ be the projector onto the negative eigenstate [67]. We must first demonstrate that $2Tr[P(\rho - \sigma)] = \|\rho - \sigma\|_1$, however this proof is identical to the classical case, (see [67]). Let us define $p = Tr(P\rho)$ and $q = Tr(P\sigma)$. Let M be a quantum operation that performs this projective measurement, so that: $M(\rho) = Tr(P\rho)|0\rangle\langle 0| + Tr[(\mathbb{I} - P)\rho]|1\rangle\langle 1|$ and

$M(\sigma) = Tr(P\sigma)|0\rangle\langle 0| + Tr[(\mathbb{I} - P)\sigma]|1\rangle\langle 1|$. Thus, we just need to follow these steps:

$$\begin{aligned} D(\rho||\sigma) &\overset{a}{\geq} D(M(\rho)||M(\sigma)), \\ &\geq \frac{4}{2}(p-q)^2, \\ &= \frac{4}{2}[2(Tr(P\rho) - Tr(P\sigma))]^2, \\ &= \frac{1}{2}\{2[TrP(\rho-\sigma)]\}^2, \\ D(\rho||\sigma) &\geq \frac{1}{2}\|\rho-\sigma\|_1^2. \ \square \end{aligned}$$

The inequality a is due to Theorem 5.43.

Theorem 5.45 (Monotonicity Under CP-maps)**.** *For any completely positive map $\Phi(\star)$, we have: $D(\Phi(\rho)||\Phi(\sigma)) \leq D(\rho||\sigma)$.*

Proof: The proof of this theorem is completely out of the scope of this introductory text.

Theorem 5.46. *Let ρ_1 and ρ_2 be density matrices with diagonal $\{p_i\}_{i=1}^n$ and $\{q_i\}_{i=1}^n$ respectively. Then $D(\rho_1||\rho_2) \geq \sum_i p_i(\log p_i - \log q_i)$.*

Proof: Let $\Phi(\star)$ be the map which annuls the off-diagonal entries of these density matrices [53]. Then, by using Theorem 5.45, we prove the theorem: $D(\rho_1||\rho_2) \geq D[\Phi(\rho_1)||\Phi(\rho_2)] = D(P||Q) = \sum_i p_i(\log p_i - \log q_i)$. $\square$.

Theorem 5.47. *Let ρ_1 and ρ_2 be density matrices with diagonal $\{p_i\}_{i=1}^n$ and $\{q_i\}_{i=1}^n$ respectively. The Von Neumann entropy of the subsystem 1 is majorized by the Shannon entropy of the diagonal entries: $S(\rho_1) \leq H(P)$.*

Proof: Choosing $\rho_2 = \frac{\mathbb{I}}{n}$ in Theorem 5.46, we have $D(\rho_1||\frac{\mathbb{I}}{n}) = \log n - S(\rho_1) \geq -\log\frac{1}{n} + \sum_i p_i \log p_i$, that is $-S(\rho_1) \geq \sum_i p_i \log p_i$, then $S(\rho_1) \leq H(P)$ [53]. $\square$

5.10 Measurements and Entropy

It is not a surprise that the behavior of the entropy after performing a quantum measurement depends on which type of measurement we performed in a quantum system. We have the two following theorems to elucidate this issue:

Theorem 5.48 (Projective Measurements Increase Entropy)**.** *Let us suppose that $\{P_i\}_{i=1}^n$ is a complete set of orthogonal projectors and let us also suppose that the state ρ, after performing the measurement is given by ρ' such that $\rho' = \sum_{i=1}^n P_i\rho P_i$. Then $S(\rho') \geq S(\rho)$ and the equality holds iff $\rho' = \rho$ [10].*

Proof: Let us apply the Corollary 5.13:

$$
\begin{aligned}
0 \leq D(\rho||\rho') &= -S(\rho) - Tr(\rho \log \rho'), \\
0 &\leq -S(\rho) - Tr[\mathbb{I}(\rho \log \rho')], \\
0 &\overset{a}{\leq} -S(\rho) - Tr[\sum_i P_i(\rho \log \rho')], \\
0 &\overset{b}{\leq} -S(\rho) - Tr[\sum_i P_i \rho \log \rho' P_i], \\
0 &\overset{c}{\leq} -S(\rho) - Tr[\sum_i P_i \rho P_i \log \rho'], \\
0 &\leq -S(\rho) - Tr[\rho' \log \rho'], \\
0 &\leq -S(\rho) + S(\rho'), \\
S(\rho) &\leq S(\rho'). \ \square
\end{aligned}
$$

Where in a we just put $\mathbb{I} = \sum_{i=1}^{n} P_i$. In b, we use the relations $P^2 = P$ and the cyclic property of the trace function. In c, we use the fact that $[P_i, \rho'] = [P_i, \log \rho'] = 0$, because $P_i \rho' = P_i \rho P_i = \rho' P_i$ [10].

Theorem 5.49 (Generalizated Measurements Can Reduce Entropy)**.** *Let us suppose that one qubit is measured using the following measurements operators* $M_1 = |0\rangle\langle 0|$ *and* $M_2 = |0\rangle\langle 1|$*, and also that the state after the measurement is unknown and it is described by* $\rho' = \frac{\sum_{i=1}^{2} M_i \rho M_i^\dagger}{Tr(\sum_{i=1}^{2} M_i \rho M_i^\dagger)}$*, then* $S(\rho') \leq S(\rho)$ [10].

Proof: Write a diagonal qubit in the basis 0, 1: $\rho = \lambda_{00}|0\rangle\langle 0| + \lambda_{11}|1\rangle\langle 1|$. Then $S(\rho) = -\lambda_{00} \log \lambda_{00} - \lambda_{11} \log \lambda_{11} > 0$ $\lambda_{00},\ \lambda_{11} \neq 0$. But $\rho' = \frac{\sum_{i=1}^{2} M_i \rho M_i^\dagger}{Tr(\sum_{i=1}^{2} M_i \rho M_i^\dagger)}$ $= |0\rangle\langle 0|\rho|0\rangle\langle 0| + |0\rangle\langle 1|\rho|1\rangle\langle 0|$ Then $(\lambda_{00} + \lambda_{11})|0\rangle\langle 0|$. After normalization $\rho' = |0\rangle\langle 0|$ Then $S(\rho') = 0 < S(\rho)$.

5.11 The Second Law of Thermodynamics - A Naive Introduction

Let us consider a quantum open system[11], *i.e.*, a quantum system with contact with some environment. We would like to observe the evolution of the quantum subsystem without paying attention to the environment. Let us also suppose that the subsystem (A) and the environment (E) are initially uncorrelated [50]. By the additivity property, we have:

$$
\begin{aligned}
\rho_{AE} &= \rho_A \otimes \rho_E, \\
S(\rho_{AE}) &= S(\rho_A) + S(\rho_E).
\end{aligned}
$$

[11] This subsection is totally inspired in [50].

We already know that the evolution of quantum open systems are described by global unitaries U_{AE}, then:

$$\rho_{AE} \Rightarrow U_{AE}\rho_{AE}U_{AE}^{\dagger},$$
$$S(U_{AE}\rho_{AE}U_{AE}^{\dagger}) \overset{a}{=} (\rho_{AE}).$$

The equality a follows because of the unitary invariance of S. Finally, it allows us to apply the sub-additivity property of $U_{AE}\rho_{AE}U_{AE}^{\dagger}$:

$$S(\rho_{AE}) = S(\rho_A) + S(\rho_E) = S(U_{AE}\rho_{AE}U_{AE}^{\dagger}) \overset{b}{\leq} S(U_{AE}\rho_A U_{AE}^{\dagger}) + S(U_{AE}\rho_E U_{AE}^{\dagger}).$$

The inequality b follows from the sub-additivity of the composite system. The equality holds *iff* the subsystem A and the environment E remain uncorrelated. If we define the total entropy as $S_{tot} \equiv S(U_{AE}\rho_A U_{AE}^{\dagger}) + S(U_{AE}\rho_E U_{AE}^{\dagger})$, then it cannot reduce. The naive assumption made in order to derive this *law* was that the subsystem and the environment were initially uncorrelated [50]. If the environment *forgets* quickly all the previous interactions with the system, then, these system–environment interactions can always be modeled as markovian processes. Hence, under this assumption, the S_{tot} will increase monotonically until attain its maximum value [50].

5.12 A Glimpse of the Quantum Entanglement

Despite the essential importance of the entanglement in quantum theory of information, this issue is not the focus of this text. We have already defined the entanglement in Chapter 3. Entanglement is a very useful resource for many interesting applications in physics such as *quantum teleportation*, *quantum cryptography* etc.

5.12.1 Pure States of a Bipartite System

Let us consider a pure state of a bipartite system. There exists cases where the reduced state of a pure state is not pure, (*e.g.*, the Bell states), then the reduced state has a non-null entropy. The question if a *pure state* $|\psi\rangle \in \mathcal{H}_1 \otimes \mathcal{H}_2$ is separable is quite easy to answer: it is enough to take the partial trace $\rho_1 = Tr_2(|\psi\rangle\langle\psi|)$ and check if $Tr(\rho_1^2) = 1$. If it is, then the state ρ_1 is pure, hence the state $|\psi\rangle\langle\psi|$ is separable. Conversely, the state $|\psi\rangle\langle\psi|$ is entangled [31]. We can define the Schmidt vector $\vec{\lambda}$, (see Schmidt Theorem in [31], Chapter 9, Section 9.2, Eq. 9.8, or in [72]), that is a k-dimensional vector, with $k \leq n = dim(\mathcal{H}_1 \otimes \mathcal{H}_2)$. The entanglement entropy is defined as the Von Neumann entropy of the *reduced* state, which is equal, by definition, to the Shannon's entropy H of the Schmidt vector (given by $\vec{\lambda}$).

Definition 5.50 (Entanglement Entropy)**.** *Let ρ_1 be the reduced state, then the Entanglement entropy is: $E(|\psi\rangle) = S(\rho_1) = H(\vec{\lambda}) = -\sum_{i=1}^{k} \lambda_i \log \lambda_i$.*

For separable states $E(|\psi\rangle) = 0$ and for maximally entangled states $E(|\psi\rangle) = \log n$ [31].

Theorem 5.51 (Nielsen's Majorization Theorem). *A given state $|\psi\rangle$ may be transformed into the state $|\phi\rangle$ by deterministic LOCC operations iff, the corresponding vectors of the Schmidt coefficients satisfy the following majorization relation: $\vec{\lambda}_\psi \prec \vec{\lambda}_\phi$.*

Proof: The proof is out of the scope of this text. But we can observe that, if we have this relation for the Schmidt coefficients, $\vec{\lambda}_\psi \prec \vec{\lambda}_\phi$, then for every Schur-convex function, we will have $H(\vec{\lambda}_\psi) \geq H(\vec{\lambda}_\phi)$. By using Theorem 5.50, we can conclude that $E(|\psi\rangle) \geq E(|\phi\rangle)$. Understanding entanglement as a resource, it is natural to think that a state can be transformed into another by means of LOCC *iff* it is more entangled than the latter.

5.12.2 Mixed States and Entanglement

Theorem 5.52 (Majorization Criterion). *If a state ρ_{AB} is separable, then the reduced states ρ_A and ρ_B satisfy the majorization relations: $\rho \prec \rho_A$ and $\rho \prec \rho_B$ [31].*

It is well known that separable states are more disordered globally than locally. In order to prove this criterion, it is sufficient to exhibit a bistochastic matrix B, such as $\vec{\lambda} = B\vec{\lambda}_A$.

Proof: Following the solution of the problem 15.4 of [31]: Let ρ be a separable state $\rho = \sum_j \lambda_j |\psi_j\rangle\langle\psi_j|$, written in its decomposition into pure product states. Then $\rho = \sum_i p_i |\phi_i^A\rangle\langle\phi_i^A| \otimes |\phi_i^B\rangle\langle\phi_i^B|$. Let us write the reduced state $\rho_A = Tr_B(\rho) = \sum_i p_i |\phi_i^A\rangle\langle\phi_i^A|$. The reduced state can be written in its eigenbasis, *i.e.*, $\rho_A = \sum_k p_k |k\rangle\langle k|$. We need to apply the Schrödinger mixture theorem (Theorem 5.32) twice. For this, we need to define two unitary matrices V and U such that: $\sqrt{p_i}|\phi_i^A\rangle = \sum_k V_{ik}\sqrt{\lambda_k^A}|k\rangle$ and $\sqrt{\lambda_j}|\psi_j\rangle = \sum_i U_{ji}\sqrt{p_i}|\phi_i^A\rangle|\phi_i^B\rangle$. If we multiply the result by its adjoint and using the orthogonality of k, *i.e.*, $\langle k'|k\rangle = \delta_{kk'}$, we will obtain $\lambda_j = \sum_k B_{jk}\lambda_k^A$. This matrix B is bistochastic. Then using the Theorem 4.7, we acquire the proof.

5.13 The Jaynes Principle in Quantum Mechanics

5.13.1 The Quantum Jaynes State

We already seen that the expectation value of an operator F_k, when the system is in a state given by ρ, is given by the rule: $\langle F_k\rangle = Tr(F_k\rho)$. When, instead of the density operator, a complete set of expected values is given, then we can define a set $\mathcal{C}$ defined by all density matrices that fulfill the conditions: $\mathcal{C} = \{\rho,\ \text{s.t. } Tr(\rho) = 1,\ Tr(\rho F_k) = \langle F_k\rangle,\ k = 1, \cdots, n\}$ [73, 74]. We should choose a state in an unbiased manner. Accordingly the Jaynes principle of the maximum entropy, the chosen state is the state that has the largest entropy still compatible with the expectations data collected from the experiments. Then $\rho_J = max_{\{\rho\in C\}}\{S(\rho)\}$. We have to maximize ρ subject to the constraints imposed by the knowledge of the expectations measured. We showed in

Theorem 5.29, that the state, which maximizes the Von Neumann entropy and simultaneously fulfills the constraint given by an expectation, is the Gibbs state $\sigma_G = \frac{e^{-H}}{Z}$, where $Tr(H\sigma_G) = \langle E \rangle$, and $Z = Tr(e^{-H})$. When we have more constraints (given by the set $\mathcal{C}$), it is easy to see that the state will be given by:

Definition 5.53 (Quantum Jaynes State)**.** *The Quantum Jaynes State ρ_J is defined as* $\rho_J = \dfrac{e^{-\sum_{k=1}^{n} \lambda_k F_k}}{Z}$, *with* $Z = Tr(e^{-\sum_{k=1}^{n} \lambda_k F_k})$.

Where the λ_k are Lagrange multipliers. As usual in Statistical Mechanics, we can calculate the expectations by performing derivatives of the partition function:

$$\langle F_k \rangle = -\frac{\partial}{\partial \lambda_k} \ln Z. \tag{5.15}$$

The vector of Lagrange multipliers can be determined, in principle, by the set $\mathcal{C}$, after performing the derivatives in the partition function Z given in Eq. 5.15. If we substitute the state ρ_J into the entropy formula, we will get: $S(\rho_J) = \ln Z + \sum_k \lambda_k F_k$. Using this last equation, we can see that $dS(\rho_J) = \sum_k \lambda_k dF_k$. Then $\lambda_k = \frac{\partial S(\rho_J)}{\partial F_k}$.

5.13.2 The Problem

The Jaynes inference scheme was studied by Buzek et al. in 1997, when they reconstructed the quantum states of 1, 2 and 3 spins from partial data [75]. A lot of similar work can be found in: [76, 77] and others. The Jaynes principle allows us to interpret the statistical mechanics as a special case of non-parametric statistical inference based on the entropic criterion [78]. This principle is the most unbiased inference scheme, since we maximize our uncertainty under the given constraints. However, is the Jaynes principle universal? In 1999, the Horodeccy gave a counterexample of fake entanglement production using the Jaynes principle in an estimate of a state based on partial information [78], and they stated the following question: What information about entanglement should be concluded, based on a given experimental mean values of an incomplete set of observables $\langle F_k \rangle = f_k$? They wanted to know if the entanglement is finally needed as a resource, then they should consider the worst case scenario, that is, should minimize entanglement under experimental constraints [79]. Put differently, the experimental values for the entanglement should be written in the form of:

$$E(f_1, \cdots, f_n) = inf_{\{i:\langle F_i \rangle_\rho = f_i, \forall i=[1,n]\}} \{E(\rho)\}. \tag{5.16}$$

Such minimization of entanglement was performed for a given mean of a Bell observable on unknown 2-qubit state [78]. Surprisingly, there were many states which achieved the minimum. Then, to reach the global minimum of the problem, the authors proposed another method of maximum entropy, based on Jaynes principle [78]. They demonstrated that the Jaynes principle, when applied to composite systems, may eventually return to states of maximum entropy compatible with the set of incomplete data which may be entangled,

even if the data comes from a separable state [78]. They suggested then that this principle should be replaced by another principle, ruled by the minimization of entanglement.

5.13.3 The Counterexample

Horodeccy considered a Bell-CHSH observable B written in the Bell basis (see Def. 3.1 and Def. 3.2). Then $B = \sqrt{2}(X \otimes X + Z \otimes Z) = 2\sqrt{2}(|\Phi^+\rangle\langle\Phi^+| - |\Psi^-\rangle\langle\Psi^-|)$, with mean value given by $0 \leq \langle B \rangle \equiv b \leq 2\sqrt{2}$. Then they applied the Jaynes inference scheme to this data (considering $\langle B \rangle = b$ as only constraint) and they obtained a quantum Jaynes state which was diagonal in the Bell basis and it was given by ρ_J.

A simple criterion of separability of Bell diagonal states is: a Bell diagonal state is separable *iff*, all its eigenvalues do not exceed $\frac{1}{2}$ [78]. Using this criterion, they showed that Jaynes principle produces fake entanglement: the Jaynes state ρ_J produced by the maximization of the entropy is inseparable, for a certain interval of values of $\langle B \rangle = b$, and they could exhibit a separable state σ compatible with the constraint given by this mean value [78].

We could think that the difference of these two inference schemes is due to the non-locality of the observable measured. But they could also exhibit a set of data which could be obtained by observers who communicate only by means of a classical channel and that still lead us toward a wrong inference scheme. Hence, if the observables exhibit some sort of correlation, Jaynes principle can fail even in the classical world.

5.13.4 An Inference Scheme: Minimization of Entanglement

The inference scheme depends on the context and, maybe, there is no way to obtain full knowledge from partial knowledge. Based on the counterexample, Horodeccy proposed an inference scheme that produces a separable state if there exists a separable state which is compatible with the constraints. Of course that an inference scheme, that does not produce fake entanglement, is a procedure that contains in one of its steps a minimization of entanglement [78]. The scheme can be stated as follows:

1. Preparation of a state ρ_B dephased in the Bell basis and compatible with the constraints.
2. This state ρ_B have the properties: $E(\rho_B) \leq E(\rho)$ and $S(\rho_B) \geq S(\rho)$.
3. Minimization of the entanglement.
4. Maximization of the Von Neumann entropy.
5. Verification if there exists a quantum separable state that is compatible with the given data.

In order to build a convex problem, Horodeccy noted that all Bell constraints[12] are linear, so they build a convex set with the states with minimal entangle-

[12] A Bell constraint is defined by $\rho \to \rho_B = \sum_{i=0}^{3} P_i \rho P_i$. We have already showed that this expansion cannot reduce entropy, see Theorem 5.48.

ment subject to the Bell constraints, because the entropy is a strictly convex function. If we maximize the entropy over a convex function we will obtain a unique state. The representative state, for the Bell constraints, is diagonal independently on the entanglement measure. Therefore, after the minimization of the entanglement in our favorite measure, we have to maximize the Von Neumann entropy and we ought to check if there exists a separable compatible with the data. [78].

5.13.5 The Solution

The example that Horodeccy gave us was very clever, but they could not state a general scheme of inference in order to completely solve the problem. Another point to consider is that they could write a convex cone with the diagonal Bell basis, by writing the state in a diagonal basis, however for more general density matrices expansions this is a much more difficult problem. They knew from the beginning that the choice of the Bell basis, (after performing the Bell dephasing step) would provide two wonderful properties: the entropy would not decrease and the entanglement would not increase. A general method of inference scheme should give us the correct convex set where to minimize the entanglement and to maximize the Von Neumann entropy.

However, an interesting question was also made by Horodeccy: "If the Jaynes state is inseparable, then the data certainly does not come from any separable state. This involves an interesting problem as well: for which type of constraints does the Jaynes scheme fail?" To our knowledge, this problem remains unsolved.

5.14 The Quantum Maximum Likelihood Principle

The statistical nature of the quantum processes is revealed in the laboratory. When an experimentalist performs repeatedly an experiment on the ensemble of identically prepared systems, he cannot control deterministically their result due to the unavoidable fluctuations [73]. The knowledge of the density matrix of a quantum system makes possible the prediction of any statistical result of any measurement performed in the system [73]. The determination of the quantum system represents an inverse problem. The inverse problem of determining the quantum state is called *quantum tomography*, see, for example [11]. This method is known, in mathematical statistics, as the Maximum Likelihood method and it was proposed by R. Fisher in the 1920's. This quantum version is identical to the classical version studied in the Section 4.7.1 and 4.7.2.

Let us consider one qubit of polarization. Assume that it is given a finite n number of copies, each in the same, but unknown quantum state which is described by the following density matrix ρ [73]. Let us consider that n photons, prepared in the same state, have been observed in m different outputs of the measurement apparatus. For one qubit of polarization, these outputs could be the vertical, horizontal, +45°, etc. Each output $|y_j\rangle\langle y_j|$, with $j = 1, \cdots, m$, has been registered n_j times with $\sum_j n_j = n$ [73]. Let us also

suppose for the sake of simplicity that this measurement is complete, *i.e.*, $H = \sum_j |y_j\rangle\langle y_j| = \mathbb{I}$. The probabilities of occurrence of various outcomes is given by the rule:

$$p_j = \langle y_j|\rho|y_j\rangle. \tag{5.17}$$

If the probabilities p_j are well-known, the problem of finding the quantum state is just a matrix inversion problem [73]. However, since only a finite number n of systems were measured, then there is no way to find out those *real* probabilities. We have in hands only the respective frequencies f_j such as $f_j = \langle y_j|\rho|y_j\rangle$ [73]. We assume that all the photons are always detected in one of the m output channels and this is repeated n times. Then we can construct the following likelihood function:

$$\mathcal{L}(\rho) = \frac{n!}{\prod_j n_j!} \prod_j \langle y_j|\rho|y_j\rangle^{n_j}. \tag{5.18}$$

Where $n_j = nf_j$. Note that Eq. 5.18 is identical to the classical formulation for the Maximum Likelihood (Section 4.7.2), where the probabilities are calculated by the Born's rule.

$$\log \mathcal{L}(\rho) = \log \frac{n!}{\prod_j n_j!} \prod_j \langle y_j|\rho|y_j\rangle^{n_j}. \tag{5.19}$$

Finding the maximum of this function (Eq. 5.19) is non-trivial and generally involves iterative methods [73, 80]. It can be shown that we need to solve the following optimization problem:

$$\begin{gathered} max_{\{\rho\in\mathcal{C}\}}\{p(j|\rho)\}, \\ \text{s.t. } Tr(\rho) = 1, \\ \rho \succeq 0. \end{gathered}$$

Where the $p(j|\rho)$ is the probability of getting the j-th outcome, given the parameter ρ. The constraints defined in the optimization problem, given by the Eq. 5.20, are known as linear matrix inequalities [81]. When the objective function $p(j|\rho)$ is linear, we have a semi-definite programming problem (SDP) [61]. Although we have efficient interior point methods for semidefinite programming, the reformulation of the problem given by Eq. 5.20 into a linear SDP form requires the introduction of auxiliary variables and constraints, that increase the dimension and therefore the difficulty of the optimization problem [81].

CHAPTER 6

Conclusions

We sincerely hope that this *invisible* parallel weaved between the Classical and Quantum Information theories has become clear. Other parallels could be evidenced if we kept following the idea developed in [31]. For instance: the issue here could follow an interesting path of the construction of a metric function for the classical space (the Fisher-Rao metric, see [31, 82]), and also for the quantum counterpart. But we focused the attention in the elements, not in the geometry of these spaces. We tried to understand a vector of probability and a quantum state as members of a classical and a quantum space of probabilities. The operations allowed in each space were then presented. In the classical case, these operations were personified by the stochastic and the bistochastic matrices. In the quantum one, we presented a brief discussion about the quantum maps and the quantum operations. We discussed some majorization properties in order to understand the *partial* ordering of these spaces. Finally, we defined the entropy function (Shannon's entropy for the classical world and the Von Neumann entropy in the quantum realm) as a measure of information and a measure of the *degree of mixing*. Some important properties of the Shannon entropy and of the Von Neumann were carefully demonstrated. The importance of the relative entropy was evidenced in both cases and some of its interesting geometrical significance were highlighted. In the quantum case, we discussed the mixture problem and the Schrödinger mixture theorem. In both cases, we displayed a short discussion on the second law of thermodynamics and we showed its connections with the Jaynes principle of maximum entropy. In the classical case we presented other inference schemes. In the quantum world we presented the Jaynes problem superficially in a short, intuitive and easy approach. This choice was justified due to the fact that this problem is still open in the quantum case.

CHAPTER 7

Appendix

Positive functionals ω are special cases of positive maps Φ. A linear map Φ is said to be positive if $f \geq 0$ implies that $\Phi(f) \geq 0$. Let us consider the following matrix $|fg\rangle = (f^*, g^*)^\dagger$. The matrix $|fg\rangle\langle fg|$ is given by:

$$|fg\rangle\langle fg| = \begin{pmatrix} ff^* & fg^* \\ gf^* & gg^* \end{pmatrix}, \tag{7.1}$$

Since $\omega(\cdot)$ is a positive linear functional, the linear map given by $\Phi(|fg\rangle\langle fg|) = (\mathbb{I} \otimes \omega)(|fg\rangle\langle fg|)$ is positive. Then:

$$(\mathbb{I} \otimes \omega)(|fg\rangle\langle fg|) = \begin{pmatrix} \omega(ff^*) & \omega(fg^*) \\ \omega(gf^*) & \omega(gg^*) \end{pmatrix} \geq 0. \tag{7.2}$$

Hence the Cauchy-Schwarz inequality follows imediately when we imply the positivity of the determinant $det[(\mathbb{I} \otimes \omega)(|fg\rangle\langle fg|)] \geq 0$, thus $|\omega(fg^*)|^2 \leq \omega(ff^*)\omega(gg^*)$ [21].

Bibliography

[1] Karl Blum. *Density Matrix Theory and Applications*. Plenum Press–New York, 1981.

[2] Claude Cohen-Tannoudji, Bernard Diu, and Frank Lalöe. *Quantum Mechanics*. Wiley, 1977.

[3] Alexander Ling et al. "Experimental polarization state tomography using optimal polarimeters". *Phys. Rev. A* **74** (2006).

[4] Michel Le Bellac. *A Short Introduction to Quantum Information and Quantum Computation*. Cambridge University Press, 2006.

[5] Grant R. Fowles. *Introduction to Modern Optics*. Dover, 1968.

[6] John D. Jackson. *Classical Electrodynamics Third Edition*. Third. Wiley, 1998.

[7] Edward L. O'Neill. *Introduction to Statistical Optics*. Addison-Wesley Publishing, 1963.

[8] Gregg Jaeger. *Quantum Information*. Springer, 2007.

[9] Henrique Di Lorenzo Pires. *Trnasformações Quânticas e Óptica Clássica*. 2007.

[10] Michael Nielsen and Isaac Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[11] Thiago O. Maciel, André T. Cesário, and Reinaldo O. Vianna. "Variational quantum tomography with incomplete information by means of semidefinite programs". *arxiv* (2011). URL: http://arxiv.org/abs/1001.1793.

[12] Fabio Benatti. *Dynamics, Information and Complexity in Quantum Systems*. Springer, 2009.

[13] Willi-Hans Steeb and Yorick Hardy. *Problems and Solutions in Quantum Computing and Quantum Information*. World Cientific, 2004.

[14] Jürgen Audretsch. *Entangled Systems: New Directions in Quantum Systems*. Wiley, 2007.

[15] Asher Peres. *Quantum theory: concepts and methods*. Kluer Academic Publishers, 1995.

[16] Scott Aaronson. *Quantum*. 2006. URL: http://www.scottaaronson.com/democritus/lec9.html.

[17] Y. Aharonov, L. Davidovich, and N. Zagury. "Quantum random walks". *Phys. Rev. A* (1993).

[18] J. Kempe. "Quantum random walks - an introductory overview". *Contemporary Physics* (2003). URL: en.scientificcommons.org/42389453.

[19] Elon Lages Lima. *Espaços Métricos*. IMPA, 1976.

[20] Marcelo O. Terra-Cunha. *Noções de Informação Quântica*. IMPA, 2007. URL: http://www.mat.ufmg.br/~tcunha/.

[21] F. Strocchi. *An Introduction to the Mathematical Structure of Quantum Mechanics – A Short Course for Mathematicians*. World Scientific Publishing Co. Pte. Ltd., 2005.

[22] Bárbara Amaral, Alexandre T. Baraviera, and Marcelo O. Terra-Cunha. *Mecânica Quântica para Matemáticos em Formação*. Impa - 28th Colóquio Brasileiro de Matemática, 2011.

[23] David Ruelle. *Statistical mechanics: rigorous results*. Imperial College Press, 1999.

[24] Matthew D. P. Daws. "Banach algebras of operators". PhD thesis. The University of Leeds - Department of Pure Mathematics, 2004. URL: ambio1.leeds.ac.uk/~mdaws/pubs/thesis.pdf.

[25] R. P. Feynman. *Statistical Mechanics - A set of Lectures*. California Institute of Technology, 1972.

[26] J. M. Steele. *The Cauchy-Schwarz Master Class - An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

[27] Thiago O. Maciel. *A Discourse on Entanglement and its Detection in Finite Dimensions*. 2011. URL: http://www13.fisica.ufmg.br/~posgrad/.

[28] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987.

[29] A. Peres. "Separability Criterion for Density Matrices". *Phys. Rev. Lett.* **77** 1413–1415 (1996). arXiv:quant-ph/9604005.

[30] Jean-Louis Basdevant and Jean Dalibard. *Quantum Mechanics*. Springer, 2002.

[31] Karol Życzkowski and Ingmar Bengtsson. *Geometry of Quantum States*. Cambridge University Press, 2006.

[32] Mateus Araújo. "Fundamentos Matemáticos da Separabilidade Quântica". Monografia - Impa. 2010. URL: http://www.mat.ufmg.br/~tcunha/MonografiaMateus.pdf.

[33] M. Junge et al. "Connes' embedding problem and Tsirelson's problem". *J. Math. Phys. 52, 012102* (2011).

[34] Tobias Fritz. "Tsirelson's problem and Kirchberg's conjecture". *Arxiv* (2011).

[35] K. Życzkowski et al. "On the volume of the set of mixed entangled states". *Phys. Rev. A* **58** 883–892 (1998). arXiv:quant-ph/9804024.

[36] K. Życzkowski and H.J. Sommers. "Hilbert-Schmidt volume of the set of mixed quantum states". *J. Phys. A: Math. Gen.* **36** 10115–10130 (2003). arXiv:quant-ph/0302197.

[37] H.J. Sommers and K. Życzkowski. "Bures volume of the set of mixed quantum states". *J. Phys. A: Math. Gen.* **36** 10083–10100 (2003). arXiv:quant-ph/0304041.

[38] M. A. Nielsen and G. Vidal. "Majorization and The Interconversion of Bipartite States". *Quantum Information and Computation* (2001). URL: http://sophia.ecm.ub.es/qic-ub/wwwsantiago/notasVidal.ps..

[39] T. Ando. "Majorization, doubly stochastic matrices, and comparison of eigenvalues". *Linear Algebra and its Applications* **118** 163 –248 (1989). URL: http://www.sciencedirect.com/science/article/pii/0024379589905806.

[40] Karol Życzkowski and Ingemar Bengtsson. "On Duality between Quantum Maps and Quantum States". *Open Systems & Information Dynamics* **11** 3–42 (2004). 10.1023/B:OPSY.0000024753.05661.c2. URL: http://dx.doi.org/10.1023/B:OPSY.0000024753.05661.c2.

[41] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1934.

[42] A. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, 1979.

[43] Charles Dunkl and Karol Życzkowski. "Volume of the set of unistochastic matrices of order 3 and the mean Jarlskog invariant". *J. of Mathematical Physics* (2009). URL: http://link.aip.org/link/?JMP/50/123521/1.

[44] Karol Życzkowski et al. "Random unistochastic matrices". *Journal of Physics A: Mathematical and General* **36** 3425 (2003). URL: http://stacks.iop.org/0305-4470/36/i=12/a=333.

[45] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley–Interscience publication, 2006.

[46] Vlatko Vedral. *Introduction to Quantum Information Science*. Oxford University Press, 2006.

[47] Claude E. Shannon. "A Mathematical Theory of Information". *The Bell System Technical J.* (1948).

[48] Annick Lesne. "Shannon entropy: a rigorous mathematical notion at the crossroads between probability, information theory, dynamical systems and statistical physics". URL: preprints.ihes.fr/2011/M/M-11-04.pdf.

[49] E. T. Jaynes. "Information Theory and Statistical Mechanics". *Phys. Rev.* (1957).

[50] John Preskill. *Lecture Notes for Physics 229: Quantum Information and Computation*. California Institute of Technology, 1998. URL: http://www.theory.caltech.edu/~preskill/ph229/.

[51] Catalin Barboianu. *Probability Guide To Gambling: The Mathematics of Dice, Slots, Roulette, Baccarat, Blackjack, Poker, Lottery and Sport Bets*. Infarom, 2008.

[52] F. Reif. *Fundamentals of Statistical and Thermal Physics*. McGraw–Hill Book Company, 1965.

[53] Dénes Petz. *Quantum Information Theory and Quantum Statistics - Theoretical and Mathematical Physics*. Springer, Berlin Heidelberg, 2008.

[54] J. Norris. *Markov chains*. Cambridge University Press, 1997.

[55] Cyrus H. A. Carzaniga. *Doubly Stochastic Converge: Uniform Sampling for Directed P2P Networks*. Tech. rep. Faculty of Informatics - Technical Report 2009/02. URL: http://old.inf.usi.ch/publications/pub.php?id=48.

[56] Jürgen Audretsch. *Entangled systems: new directions in quantum physics*. Wiley, 2006.

[57] Mani R. Gupta. "An Information Theory Approach to Supervised Learning". PhD thesis. Stanford University, 2003. URL: www.ee.washington.edu/research/guptalab/publications/thesis.pdf.

[58] Thomas M. Cover. *Physical Origins of Time Asymmetry–Chapter 5*. Cambridge University Press, 1994.

[59] Peter E. Blöchl. *Theoretical Physics IV - Statistical Physics*. Clausthal Zellerfeld - Germany, 2000 - 2011. URL: http://orion.pt.tu-clausthal.de/atp/phisx.html.

[60] E. T. Jaynes. "Information Theory and Statistical Mechanics. II". *Phys. Rev.* (1957).

[61] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL: http://www.stanford.edu/~boyd/cvxbook/.

[62] Sir R. A. Fisher. "On an absolute criterion for fitting frequency curves". *Messenger of Mathematics* (1912). URL: http://www.jstor.org/stable/2246266.

[63] Sir R. A. Fisher. "On the Mathematical Foundations of Theoretical Statistic". *Phil. Trans. R. Soc. Lond. A* (1922).

[64] John Aldrich. "R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922". *Statistical Science* (1997). URL: http://projecteuclid.org/euclid.ss/1030037906.

[65] Shmuel Schreiber. "On a Result of S. Sherman Concerning Doubly Stochastic Matrices". *Proc. Am. Math. Soc.* (1958). URL: http://www.jstor.org/pss/2032985.

[66] Alfred Wehrl. "General Properties of Entropy". *Rev. Mod. Phys.* **50** (1978).

[67] Mark M. Wilde. *Quantum Information and Entropy*. 2011. URL: http://www.markwilde.com/teaching/notes/.

[68] Herbert B. Callen. *Thermodynamics and an introduction to thermostatistics*. John Wiley & Sons, 1985.

[69] Dénez Petz. *From thermodynamics to quantum theory. Part I: Equilibrium.* 2001. URL: http://www.renyi.hu/~petz/lessrecent.html.

[70] Mary Beth Ruskai. "Inequalities for quantum entropy: A review with conditions for equality". *Journal of Mathematical Physics* (2002). URL: http://link.aip.org/link/?JMP/43/4358/1.

[71] Eric Carlen. "Trace Inequalities and Quantum Entropy: An introductory course". Lecture Notes. 2009. URL: www.mathphys.org/AZschool/material/AZ09-carlen.pdf.

[72] Fernando Iemini De Rezende Aguiar. *Emaranhamento em Sistemas de Partículas Indistinguíveis*. 2011. URL: http://www13.fisica.ufmg.br/~posgrad/.

[73] M. G. A. Paris and J. Rehacek. *Quantum State Estimation*. Springer - Berlin Heidelberg, 2004.

[74] E. T. Jaynes. "Information Theory and Statistical Mechanics". *Statistical Physics* (1963).

[75] V. Buzek et al. "Reconstruction of quantum states of spin systems via the Jaynes principle of maxim um entropy". *Journal of Modern Optics* (1997). URL: www.quniverse.sk/rcqi/mypapers/97jmo2607.pdf.

[76] Gabriel Drobny and Vladimir Buzek. "Reconstruction of motional states of neutral atoms via maximum entropy principle". *Phys. Rev. A* (2002).

[77] A. K. Rajagopal. "Quantum entanglement and the maximum-entropy states from the Jaynes principle". *Phys. Rev. A* (1999).

[78] Ryszard Horodecki, Michał Horodecki, and Paweł Horodecki. "Entanglement processing and statistical inference: The Jaynes principle can produce fake entanglement". *Phys. Rev. A* (1999).

[79] Ryszard Horodecki et al. "Quantum entanglement". *Rev. Mod. Phys.* (2009).

[80] A. I. Lvovsky. "Iterative maximum-likelihood reconstruction in quantum homodyne tomography". *Quantum and Semiclassical Optics* (2004). URL: http://www.citebase.org/abstract?id=oai:arXiv.org:quant-ph/0311097.

[81] Douglas S. Gonçalves et al. "Local solutions of Maximum Likelihood Estimation in Quantum State Tomography". *arXiv: quant-ph* (2011).

[82] S. I. Amari et al. *Differential Geometry in Statistical Inference*. Institute of Mathematical Statistics, 1987.