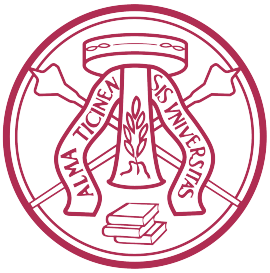


UNIVERSITÀ
DI PAVIA

Dipartimento di Fisica

Dottorato di ricerca in fisica XXXV ciclo

Towards a generalised Shannon theory: information content of a source

---

Leonardo Vaglini

Advisor:
Prof. Paolo Perinotti

submitted to the graduate school in physics in partial fulfillment of the requirements for the degree of
Dottore di ricerca in fisica (Doctor of Philosophy in Physics)
at the university of pavia

Contents

CHAPTER 1

Introduction

In the last century, the relationship between physics and information has become more and more tight. A famous example is inspired by the works of several authors—Jakob Bekenstein, Stephen Hawking and Leonard Susskind among others—concerning the thermodynamics of black holes and the entropy of gravitational systems [1, 2, 3, 4, 5, 6, 7, 8]. In particular, they discovered that the entropy of a black hole is proportional to the horizon area of the black hole itself. Elaborating on this, and interpreting the entropy as a measure of the number of degrees of freedom needed to describe the physics in a given region, it was concluded that this number is bounded by the area that circumscribes the region itself. This is, roughly speaking, the essential content of the so-called *holographic principle* [9, 8], a property that many physicists believe should be a tenet of a quantum theory of gravity (see [10] for a comprehensive review). A phenomenon that is completely analogous to the latter is that of *area laws* in many-body systems. A prototypical example of a many-body system is a lattice, whose evolution is described by a local Hamiltonian, and the whole lattice can be thought of as being in a pure state. In such a system one can define the entropy of a subregion of the lattice, which measures the entanglement between such a region and the rest of the lattice, and it is accordingly called the *entanglement entropy*. What is found is that such entropy tipically scales with the area of the boundary of the subregion, instead of scaling with the volume, and such scaling laws are of interest for the study of critical behaviour in complex quantum systems and for improving the efficiency of numerical algorithms in quantum many-body physics (see [11] and references therein). These two examples show that the interplay between physics and information—the

latter being represented by a suitable entropic quantity—can be very fruitful for a deeper understanding of fundamental physical phenomena.

The strong belief in the fact that capitalising "on the findings and outlooks of information theory" may definetely reshape our understanding of physics was put forward—and epitomised in the motto *it from bit*—by John Archibad Wheeler [12]. Such a phrase summarises a thought-provoking proposal for a new paradigm to be pursued in physics. In his words:

> "[...]every it—every particle, every field of force, even the space-time continuum itself— derives its function, its meaning, its very existence entirely—even in some contexts indirectly— from the apparatus-elicited answers to yes-or-no questions, binary choices, *bits*."

A partial fulfillment of the proposal made by Wheeler is represented by the denouement of the program, initiated by Birkhoff and von Neumann[13], of an axiomatisation of Quantum Theory (QT) that makes no reference to the mathematical structure of Hilbert spaces or C*-algebras, but relies on principles that refer to operational and informational statements only. The path of the operational approach, where the building blocks consist of prepararation and measurement apparatuses, and postulates concern how these combine to give outcome probabilities, was taken by Ludwig [14], who however failed to completely reconstruct the Hilbert space structure from purely operational principles. With the development of quantum information theory in the 80s and 90s it was understood that the mathematics of QT entails startling properties, such as no-cloning and quantum teleportation, and this led to a question: is it possible to reverse the logic and derive QT from its informational features? This question gave rise to a new research programme which was condensed in a sentence "quantum foundations in the light of quantum information" [15], which, in turn, gave a positive answer [16, 17]

The principles that are adopted in the operational frameworks in which QT is derived do not contain references to a measure of information. However, to study statements as the aforementioned holographic principle we cannot steer clear of defining what we mean by *information content* of a physical system. The work presented in this thesis stems from this consideration, and investigates how, and in which conditions, it is possible to give a proper definition of information content in the framework of Operational Probabilistic Theories (OPTs).

Now, more often than not, in physics the word *information* is associated with the word *entropy*. Historically, the latter became part of the physicists' language in 19th century, introduced by the German physicist and mathematician Rudolf Clausius, while he was developing the second law of thermodynamics between 1850 and 1870. It was not too long before Ludwig Boltzmann, while laying the foundations of statistical mechanics, defined the entropy as the logarithm of the number of microstates that correspond to a given macrostate.

Then, Willard Gibbs refined the definition of Boltzmann, furnishing a mathematical formula that quantifies, to some extent, the *lack of knowledge* due to the 'coarse-graining' of microstates, which in turn defines the thermodynamical macrostate[1]. Around seventy years later, the same formula was proposed by Claude Eldwood Shannon, an american mathematician and engineer, in order to quantify the uncertainty, measured in bits[2], associated with a source of information represented by a classical system. About the name he decided to give to this quantity, the following quotation is credited to Shannon [18]:

> "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. 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, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' "

and indeed, that quantity, that shares the same mathematical expression with the Gibbs entropy, is nowadays called the *Shannon entropy* (usually denoted as H).

Shannon was well aware of the Gibbs formula from statistical mechanics; his breakthrough was to recognise that the same formula fully characterises the statistical nature of an information source. On the one hand, considering an experiment whose outcomes can occur according to a probability distribution, he posed the question: how much *uncertainty* do we have about the outcome? The first justification he provided relies on an axiomatic approach, where he defines a set of postulates that a reasonable measure of uncertainty should satisfy (see the original work [19], and [20]). On the other hand, he also gave a characterisation that is *fully operational*, in the form of a theorem that is commonly known as *first Shannon theorem* or (classical) *noiseless coding theorem*. This establishes the Shannon function as a measure of the *information content* of a classical *source of information*. The simplest description of a classical information source is in terms of a sequence $X_1, \ldots, X_N, \ldots$ of random variables that are identically and independently distributed, so that they all take values in the same finite set of symbols $S = \{x_i\}$. The source then produces messages $x_{i_1} \ldots x_{i_N}$ of variable length N. Let us further simplify the discussion by assuming that $S = \{0, 1\}$, namely that the messages are represented by strings of bits. Roughly speaking, the theorem states that the smallest factor by which the messages can be shortened without losing those ones that are most likely

[1] Recall that, if $\{p_i\}$ is the probability distribution associated with the microstates, the Gibbs entropy is given by $S_{\mathrm{G}} = -k_{\mathrm{B}} \sum p_i \log_2 p_i$, where k_{B} is the Boltzmann's constant.

[2] *bit*, the portmanteau of binary digit, is introduced in the literature by Shannon himself, who, however, credited the coining to John Wilder Tukey, an american mathematician.

to occur, is exactly the Shannon entropy H associated with the source. In this sense, the theorem answers a fundamental question: how much resources do we need in order to faithfully convey the messages of a source? Or, in other words, to what extent can we get rid of redundancies in the messages, without losing pieces of information? This kind of problem is commonly known in the field of information theory as *compression task*. Needless to say that this issue is of practical relevance, but its solution also shed a new light on what it means to measure information from a foundational viewpoint. All that being said, both names—uncertainty and information content—may well be applied to the Shannon function, given the results that he proved in his work.

The problem of quantum information compression was firstly solved by a theoretical physicist, Benjamin Schumacher[3], in 1995 [21]. The idea was to translate all the elements of the classical problems in quantum terms, "instead of simply applying classical information theory to probabilities derived from quantum rules". Then: i) a source of information becomes an ensemble of arbitrary pure states lying in a given Hilbert space, and giving rise to a signal that can be described by a mixed state; ii) the messages, generalising the i.i.d. classical setting, are factorised states of the form $\rho^{\otimes N}$, processed with quantum channels ; iii) last but not least, the elementary unit of information, the analogous of the classical bit, is the "quantum bit" or *qubit*[4], namely a 2-level quantum system, such as the spin of the electron. The role of the Shannon entropy is then played by the von Neumann entropy S, introduced by John von Neumann as a generalisation of the Gibbs entropy[22]. This entropy answers a slight variation of the question we posed before: to what extent can we get rid of redundancies in the messages, without losing *quantum* pieces of information? The answer is that if we have a message made of N qubits from a source described by a density matrix ρ, then it cannot be faithfully transmitted using less than $S(\rho)$ qubits. This provides the von Neumann entropy with the same operational interpretation enjoyed by the Shannon entropy in classical theory: it is a measure of the amount of quantum information, measured in qubits. Furthermore, Schumacher gave a profound insight, showing that in order to achieve a reliable transmission of quantum information, one cannot neglect quantum correlations, in a motto: *preserving quantum information is equivalent to preserving entanglement.*

Now, the question the present thesis work hinges upon is how to define a proper measure of information in a context in which the features of Classical Theory (CT) and QT are not assumed from the beginning. The framework adopted here is that of Operational Probabilistic Theories, that has proven to be fruitful for the investigation of quantum features from a broader point of view. There are three notions of entropy that have been introduced in the literature, defined in terms of classical information quantities as follows: i)

[3]Curiously, notice that, as Schumacher also mentioned in [21], he has been a Ph.D. student of Wheeler.

[4]The term, as reported by Schumacher himself, was coined in jest during a conversation between him and Wootters.

The *measurement entropy* of a system, namely the infimum Shannon entropy of any possible measurement on the system, quantifies the minimum measurement uncertainty, provided that the system is prepared in the state of interest. ii) The *decomposition (or mixing) entropy*, namely the infimum of the Shannon entropies over all possible ways of preparing the system's state as a mixture of pure states, quantifies the minimum uncertainty for a preparation of a state with respect to pure states. iii) The supremum of the Shannon mutual information between two random variables related respectively to measurements on the system and decompositions of the state of interest, quantifies the maximum *accessible information*. All of these generalisations of the Shannon and von Neumann entropies are obtained as the result of an approach that may be considered an hybrid between operational and axiomatic. On the one hand, as it is also clear by the above description, they have an operational meaning, since they refer to preparations, measurements or both. On the other hand they rely on a quantity that has its full justification as a measure of information in a classical context, therefore their understanding as an uncertainty measure is ultimately inherited from the interpretation of the Shannon function as such.

In the present dissertation we pursue a fully operational approach to generalise the notion of information content to probabilistic theories that do not necessarily satisfy the laws of classical or quantum physics. In particular, we propose a definition that stems from the insights of the classical and quantum noiseless coding theorems. In other words, by reversing the logic, we introduce an operational entropy as the minimal compression rate of a suitable generalisation of the task mentioned above in the classical and quantum scenarios. The main assumption, that guarantees the meaningfulness of our definition, is that of "digitisability". As we saw, classical and quantum information content have their own unit of measure, the bit and qubit respectively. In general, there can be probabilistic theories that do not have the equivalent of these systems, whereby it might be not possible to define the compression rate unambiguosly. Precisely, our assumption consists of: i) the supposed existence of *at least* one system, say B, such that any other can be encoded on a suitable array made of a finite number of B systems; ii) if more than one of such systems exist then they are equivalent. It is worth emphasising that this assumption is not so limiting, in light of the fact that all the probabilistic theories known in the literature satisfy it. This work unifies the classical and the quantum results in a general setting, and consents to deal with problems and statements related to the information content via a top-down approach. In particular, one can investigate its general properties, regardless of the characteristic traits of classical and quantum theories, and study the behaviour of the information content in specific theories to establish links between properties of the theory and the general features of the information content.

In **chapter 2** we review in detail the classical and quantum noiseless coding theorems, since our scope is to extrapolate the operational meaning that such results confer to the Shannon and von Neumann entropies respectively. In

order to do this, we give a summary of the mathematical structure of CT and QT, and this will also serve to fix the notation. We will also remind the reader of the notions of classical and quantum weak typicality, since they will be used in chapter 6 and chapter 5 respectively.

In **chapter 3** we provide an extended summary of the OPT framework, with an account of the operational and probabilistic language. We will introduce two properties that are relevant in our dissertation: *conditioning* (or *strong causality*), and *steering*. We then provide a brief recap of fermionic information theory, that will be treated in chapter 5, and of Bilocal Classical Theory, considered in chapter 6. We finally review the definition and the main properties of the three different definitions of entropy already introduced in the literature, along with a generalisation of fidelity.

In **chapter 4** we introduce the notion of information content, extending to the OPT framework the compression task. Along the path we pay the due attention to generalise all the basic elements of a communication scheme and the figure of merit that assesses the error in absence of the characteristic features of CT and QT. We illustrate the "digitisability" assumption, that delimits the class of theories in which our definition applies. Subsequently, we analyse the impact of some additional assumptions on the definition, and derive simplified expressions that are of great help when dealing with specific theories, showing that in the classical and quantum case the information content exactly reduces to the Shannon and von Neumann entropies respectively. We then investigate some properties of the information content, such as subadditivity and invariance under reversible transformations. We also show that, in general, the information content is not a measure of purity for state: while it is generally true that a state with vanishing information content is pure, the viceversa is proved to be false. In particular, if the compositional law of the theory at hand does not preserve purity, then there must exist pure states with strictly positive information content. In proving this, we see that the latter is bounded from below by one of the entropy introduced in the literature, the optimised accessible information.

In **chapter 5** we apply the machinery developed in chaper 4 to the case of Fermionic Theory (FT). This is a case of particular interest, since it is an example of a theory that does not satisfy the local process tomography property. Indeed, due to the parity superselection rule that forbids superpositions of vectors of different parity, the theory is strictly bilocal tomographic, and this means that transformations that are locally indistiguishable can behave in a very dissimilar way when applied locally on enlarged states. Therefore, it is by no means obvious that the information content has to match the fermionic von Neumann entropy of the fermionic source, represented by a state with definite parity. Nonetheless, relying on the Jordan Wigner isomorphism that maps fermionic systems on a suitable number of qubit systems, we are able to prove that no differences show up with respect to the quantum case.

In **chapter 6** we finally consider a toy model, named Bilocal Classical The-

ory (BCT). This theory is also strictly bilocal, but the feature that makes it rather peculiar is that independent preparations in pure states give rise to a mixture when composed. This property lies at the basis of the differences between the information content and all the other entropies. This theory is an explixit counterexample to the conjecture that one of the entropic functions proposed so far in the literature satisfies a generalised noiseless coding theorem. Remarkably, this happens in a theory where all of them boil down to the same quantity, the Shannon entropy of the state that describes the information source. Nevertheless, it might be true that in the presence of additional hypotheses, at least one of the entropies may represent the minimal compression rate in a coding theorem.

We then conclude the thesis drawing our conclusions in **chapter 7** and discussing possible directions for future research.

1.1 Symbols and acronyms

In the following, some symbols and acronyms used throughout the work are listed:

- $\mathcal{H}, \mathcal{K}, \ldots$ Hilbert space
- $\mathcal{L}(\mathcal{H}_\mathrm{B}, \mathcal{H}_\mathrm{A})$ set of linear operators from $\mathcal{H}_\mathrm{B}$ to $\mathcal{H}_\mathrm{A}$
- $\mathcal{L}(\mathcal{H}_\mathrm{A})$ set of linear operators on $\mathcal{H}_\mathrm{A}$
- $\mathcal{L}(\mathcal{H}_\mathrm{A})_>$ set of positive linear operators on $\mathcal{H}_\mathrm{A}$
- $\sigma^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\sigma^y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$, $\sigma^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ Pauli matrices
- $\mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots$ systems of a probabilistic theory
- $\mathrm{N_F}$ system corresponding to N Local Fermionic Modes
- $\mathsf{St}(\mathrm{A})$ set of states of the system A
- $\mathsf{Eff}(\mathrm{A})$ set of Effects of the system A
- $\mathsf{St}_\mathbb{R}(\mathrm{A})$ linear span of the set of states $\mathsf{St}(\mathrm{A})$
- $\mathsf{Eff}_\mathbb{R}(\mathrm{A})$ linear span of the set of effects $\mathsf{Eff}(\mathrm{A})$
- D_A size of the system A
- $\mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ set of transformations from system A to system B
- $H(\mathbf{p})$ Shannon entropy of the probability distribution $\mathbf{p} := \{p_i\}$
- $S(\rho)$ von Neumann entropy of the density matrix ρ
- $T_\delta^N(\mathbf{p})$ (N, δ)-*typical set* associated with $\mathbf{p}$
- $\mathsf{H}_\delta^N(\rho)$ (N, δ)-typical subspace associated with the quantum state ρ
- BCT Bilocal Classical Theory
- CT Classical Theory
- FT Fermionic Theory
- GPT Generalised Probabilistic Theories
- LFM Local Fermionic Mode
- OPT Operational Probabilistic Theories
- POVM Positive Operator-Valued Measures
- QT Quantum Theory

CHAPTER 2

Noiseless coding in classical and quantum information theories

In 1948, with his seminal paper, Shannon gave a solid mathematical foundation to the classical theory of information and communication [19]. In particular, he made clear the link between the notion of entropy and the information content of a classical source via the so called *first Shannon theorem*. This result justifies our comprehension of the entropy as a quantifier of uncertainty about the outcome of a classical probabilistic experiment, in terms of an operational task: information compression. More precisely, the theorem establishes that the Shannon entropy associated with the source is the minimal number of bits per length of the received message that are necessary in order to reliably compress the source, in the limit of long messages.

In order to prove this result, Shannon considered a block-coding strategy, allowing for an error in the decoding of the compressed messages. This strategy works in the asymptotic regime thanks to the powerful notion of typicality, developed by Shannon himself. Imagine a source that emits messages in the form of sequences, where each letter of the message is drawn from a finite set of allowed symbols. For fixed length of the sequences, say N, the idea was to recognize that there exist proper subsets of the set of all possible sequences, say T_N, that, for very long messages, have two really nice properties: i) they contain those sequences that are more likely to be emitted (whereby they are called typical), and ii) they have cardinality exponentially smaller than T_N. The key point is that the cardinality of the typical sets and the probabilites of typical sequences solely depend on the Shannon entropic function, there-

fore, the latter fully characterizes the statistical nature of the source. The coding strategy for a given length N is then to store all the typical sequences, and throwing away all the rest, and by taking N sufficiently large, the error probability can be made arbitrarily small.

The analogous problem of *quantum* noiseless coding was succesfully solved in 1995 for the first time, by B. Schumacher in [21] (with further refinements considered in [23, 24]). The question he addressed is how much *quantum* information can be compressed, measured in qubits per symbol. The answer turns out to be quite simple and is, in form, essentially equal to the classical case: the minimal rate that allows for ideal compression of quantum information is equal to the von Neumann entropy of the density matrix that represents the source. Therefore, the latter can be interpreted as the information content of a quantum source of information, measured in qubits. On the one hand, Schumacher considered an ensemble of pure states, not necessarily orthogonal, as a quantum souce, which gives rise to a signal represented by a mixed state; the task is to transpose the signal faithfully, minimising the average error. But in quantum theory, a mixed state can also be interpreted as the marginal state of a system that is entangled with a reference one, with the whole system being in a pure state. The insightful observation of Schumacher is that $S(\rho)$ is the ultimate compression rate with which we can transmit the local messages, while preserving quantum entanglement with a reference system. This latter fact makes the popular motto *transmitting quantum information means preserving entanglement* quite solid.

Therefore, the classical and quantum noiseless coding theorems give us two useful insights: i) measuring how much information is carried, or contained, in a physical system, is equivalent to asking how much redundancy we can eliminate in a compression task; ii) preserving information does not only mean reproducing the local messages, but also taking care of the correlations with an environment.

In this chapter we thoroughly review the aforementioned results, since they constitute the ground for our operational definition of information content. Along the way, we give an overview of the mathematical structure of both classical and quantum theory, with the purpose of fixing the notation and the terminology that will be used throughout the work when we refer to such theories. We will also remind the reader of the notions of classical and quantum weak typicality, since they will be useful also in chapter6 and chapter 5 respectively.

2.1 Classical Shannon theory

The aim of this first section is to give a self-consistent review of the first Shannon theorem in its simplest form. In particular, we briefly review the setting in which the noiseless coding theorem is proved. To this end, we will recall the mathematical structure of classical theory and the definition of Shannon

entropy, along with some related properties. We then remind the reader of the notion of weak typicality, a powerful tool that also finds application in the quantum setting, as we will see later, and finally we conclude with the statement of the celebrated Shannon's noiseless coding theorem.

2.1.1 Summary of Classical Theory

Classical information theory regards processing of information that is stored in *classical* systems by means of *classical* dynamics. Therefore, we must first recall the basic facts about classical theory, namely which are the mathematical representatives of classical states, transformations and measurements. This overview has the main scope of fixing the notation when dealing with results concerning CT throughout the work.

A *type* of classical *system* is associated with real vector spaces $\mathbb{R}^{d_A}$, and different types are associated with different values of the dimension d_A. A classical system is associated with a given set of symbols, X, whose cardinality $|\mathsf{X}|$ equals the dimension d_{A}[1]. The set of states of a given system A, which is denoted by $\mathsf{St}(\mathrm{A})$, is made of sub-stochastic vectors in these spaces, namely by vectors $\mathbf{p}$ satisfying

$$\|\mathbf{p}\| \coloneqq \sum_{i=1}^{d_A} |p_i| \leq 1. \tag{2.1}$$

This norm $\|\cdot\|$ induces the following distance on the set of classical states

$$\|\mathbf{p}-\mathbf{q}\|_1 \coloneqq \frac{1}{2}\|\mathbf{p}-\mathbf{q}\| = \frac{1}{2}\sum_{i=1}^{d_{\mathrm{A}}} |p_i - q_i| \tag{2.2}$$

Notice that, in particular, pure states are represented by the canonical basis vectors $\mathbf{e}_i$ of $\mathbb{R}^{d_{\mathrm{A}}}$, that are those vectors with all vanishing components except for the i-th one. Therefore the linear span of the set of states coincide with the vector space $\mathbb{R}^{d_{\mathrm{A}}}$, namely

$$\mathsf{St}_{\mathbb{R}}(\mathrm{A}) := \mathsf{Span}_{\mathbb{R}}[\mathsf{St}(\mathrm{A})] = \mathbb{R}^{d_{\mathrm{A}}} \tag{2.3}$$

Moreover, we also observe that the set of states of any classical system is a simplex, that is the convex hull of a finite set of affinely independent points. The set of those states such that $\mathbf{p} = \{p_i\}_{i=1}^{d_{\mathrm{A}}}$ is a probability distribution also called the set of *deterministic* states, will be denoted by $\mathsf{St}_1(\mathrm{A})$. Another useful way in which we can think of a deterministic state, is as a random variable X, which is fully specified by an alphabet $\mathsf{X} = \{x_1, \ldots, x_{d_{\mathrm{A}}}\}$ (also denoted as $\mathsf{Rng}(X)$ and called *range*) on which it takes values, and a probability distribution $\{p(x_i)\}_{i=1}^{d_{\mathrm{A}}}$.

[1] Clearly, different systems of the same type are associated with a different set of symbols that have the same cardinality, but they are completely equivalent from an informational standpoint.

A measurement is represented by a set $\{\mathbf{x}_j\}_{j\in\mathsf{Y}}$ of unit-dominated positive vectors, i.e. those $\mathbf{x} \in \mathbb{R}^{d_\mathrm{A}}$ such that $0 \leq x_k \leq 1$ for any $k = 0, \ldots, d_A$, satisfying the additional condition

$$\sum_{j\in\mathsf{X}} \mathbf{x}_j = (1, \ldots, 1) =: \mathbf{e}.$$

The elements of the measurement $\mathbf{x}_j$ are also called *effects* and the vector $\mathbf{e}$, that has all components equal to 1, is called the determinsitic effect. When an effect $\mathbf{x}_j$ of a measurement $\{\mathbf{x}_j\}_{j\in\mathsf{Y}}$ is applied to a state $\mathbf{p}$, the probability of the outcome is the scalar product of the two vectors $\mathbf{x}_j \cdot \mathbf{p}$. Notice that pure states can be perfectly discriminated via the measurement whose effects are, in fact, the canonical basis vectors.

Transformations from system A to system B are represented by $d^\mathrm{B} \times d^\mathrm{A}$ sub-stochastic matrices (i.e. matrices whose columns are a sub-stochastic vector) $\mathbf{M}$ acting on the probability vectors by multiplication $\mathbf{p} \to \mathbf{Mp}$; deterministic transformations are stochastic matrices (namely, whose columns are normalized probability distributions) while reversible ones are permutations.

The action of two transformations in sequence is simply given by matrix multiplication while the composition of two systems A and B is obtained with the tensor product, where the pure states are the vectors of the form $\mathbf{e}_i \otimes \mathbf{e}_j$. Notice that the Krein-Milman theorem implies that any state of the joint system AB, say $\mathbf{\Pi}$, admits a unique pure state decomposition of the form

$$\mathbf{\Pi} = \sum_i \Pi_{ij} \mathbf{e}_i \otimes \mathbf{e}_j, \tag{2.4}$$

where $\sum_{ij} \Pi_{ij} \leq 1$.

2.1.2 A measure of uncertainty: the Shannon entropy

The main role in the field of information theory is played by the so called Shannon entropy [19]. Its introduction in the information-theoretic context responds to the urgency of a quantification of uncertainty about a classical experiment that we are going to perform. Complementarily, such a quantity can also be interpreted as the average information gain that is obtained once the experiment is performed, and we have learnt the outcome, say i. Roughly speaking, the quesiton is: if we have a classical system described by a probability distribution $\mathbf{p} := \{p_i\}_{i\in\mathsf{X}}$, where each $i \in \mathsf{X}$ can be think of as associated with an event that may occurr, what is our uncertainty about the outcome? The answer is provided, indeed, by the Shannon function, whose definition is recalled in the following.

Definition 2.1.1. *Let* $\mathbf{p} = \{p_i\}_{i\in\mathsf{X}}$ *be a probability distribution associated with a state of a classical system* A *with dimension* d_A. *The Shannon entropy* $H(\mathbf{p})$

is defined as

$$H(\mathbf{p}) := -\sum_{i=1}^{d_{\mathrm{A}}} p_i \log p_i, \tag{2.5}$$

with the convention that $0 \log_2 0 \equiv 0$. *Sometimes we will simply denote it by* $H(\mathrm{A})$, *without making explicit reference to the state in which the system* A *is whenever it is clear from the context.*

Remark 2.1.1. In his original work, Shannon also gave a heuristic motivation that justifies the mathematical function $H(\mathbf{p})$ that should describe such a quantity, formalised in terms of three "reasonable" assumptions:

1. $H(\mathbf{p})$ is a continuous function of its variables;

2. if $\{p_i\}_{i\in\mathsf{X}}$ is the flat distributions, i.e. $p_i = \frac{1}{|X|}$, then H should be a monotone increasing function of $|X|$. This is the obvious intuition that the greater the choice, the greater the uncertainty;

3. "If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H."

The third is directly quoted from [19], and it can be rephrased in a mathematical fashion as follows: consider a probability distribution $\{p_i\}_{i\in\mathsf{X}}$ such that there exist a partitioning $\{\mathsf{X}_j\}_{j\in\mathsf{Y}}$ of X, a probability distribution $\{q_j\}_{j\in\mathsf{Y}}$ and, for any $j \in \mathsf{Y}$, a probability distribution $\{q_{i|j}\}_{i\in\mathsf{X}_j}$ such that $p_i = q_j q_{i|j}$ for any $i \in \mathsf{X}_j$ and for any $j \in \mathsf{Y}$ (see also figure 2.4). Then, assumption 3 tells us that for any such distribution one has that $H(\{p_i\}_{i\in\mathsf{X}})$ satisfies

$$H(\{p_i\}) = H(\{q_j\}) + \sum_{j\in\mathsf{Y}} q_j H(\{q_{i|j}\}).$$

With the aforementioned hypotheses, Shannon showed that $H(\mathbf{p})$ must necessarily be of the form $-K\sum_i p_i \log p_i$, with K a constant (we refer the reader to his work for the proof, see also the famous work by rényi [20]). Therefore, setting $K = 1$, which corresponds to measure the amount of uncertainty in terms of bits, one is led to the definition 2.1.1.

Shannon was aware of the fact that similar formulas hold for the thermodynamical entropy in statistical mechanics. His intuition resides in recognising its central role in quantifying the amount of information of a classical information source, not only via the foregoing heuristic description, but also through the noiseless coding theorem, one of the main result that he presented in [19].

Based on Shannon function, one can build other information-theoretic quantities. Consider, for instance a joint classical system AB in a state $\mathbf{\Pi} = \sum_i \Pi_{ij} \mathbf{e}_i \otimes \mathbf{e}_j \in \mathsf{St}_1(\mathrm{AB})$. The *joint Shannon entropy of* A *and* B, which describes our lack of knowledge about the pair (i, j), is then obtained in terms

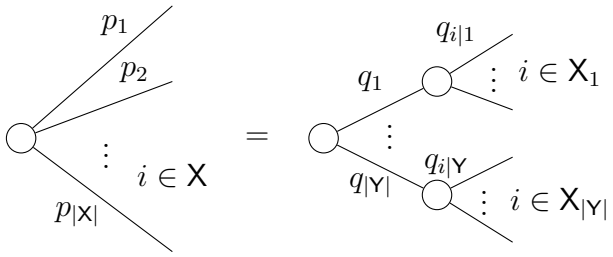


Figure 2.1: Here is a pictorial representation of what is described in remark 2.1.1. On the left side, only one choice is made according to the probability $\{p_i\}_{i\in\mathsf{X}}$. On the right hand side, the second choice is *conditioned* on the first outcome $j \in \mathsf{Y}$, and it is made according to the probability distribution $\{q_{i|j}\}_{i\in\mathsf{X}_j}$ on a set possible outcomes that is a subset X_j of X. The two-step experiment, that is a conditional test, must give rise to same probabilities for the whole set of outcomes X as the one on the left side.

of the probability distribution $\{\Pi_{ij}\}$ as follows

$$H(\mathrm{AB}) := -\sum_{i,j=1}^{d_\mathrm{A},d_\mathrm{B}} \Pi_{ij} \log \Pi_{ij}. \tag{2.6}$$

Clearly, this can be readily extended to the case of N systems in the obvious way. Along with this one can also consider the entropy $H(\mathrm{A})$ and $H(\mathrm{B})$ of the marginal states on A and B, that are given in terms of the probability distributions $\{p_i = \sum_{j=1}^{d_\mathrm{B}} \Pi_{ij}\}_{i=1}^{d_\mathrm{A}}$ and $\{q_i = \sum_{i=1}^{d_\mathrm{A}} \Pi_{ij}\}_{j=1}^{d_\mathrm{B}}$ respectively via the formula (2.5). Now, imagine that we have learnt the value of j. The last entropic quantity that we can construct with the notions encountered up until now is the mutual information. Again, consider a joint system AB in a state $\Pi \in \mathsf{St}_1(\mathrm{AB})$ and the sum of the entropies of the marginal states $H(\mathrm{A})$ and $H(\mathrm{B})$. It might be the case that in this way we are overcounting some information that is in common between A and B. If we then subtract the joint entropy $H(\mathrm{AB})$ from this sum we are left with with this shared piece of information, that is the *mutual information of* A *and* B

$$H(\mathrm{A} : \mathrm{B}) := H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{AB}). \tag{2.7}$$

The quantities that we have defined so far have several desirable properties that we list in the following theorem

Theorem 2.1.1 (Properties of the Shannon entropy)**.** *With the notation established above, the following hold:*

1. *the joint entropy and the mutual information are symmetric in their arguments, namely* $H(\mathrm{A} : \mathrm{B}) = H(\mathrm{B} : \mathrm{A})$ *and* $H(\mathrm{AB}) = H(\mathrm{BA})$*;*

2. *(subadditivity) for any* A *and* B *and any* $\Pi \in \mathsf{St}_1(\mathrm{AB})$ *one has* $H(\mathrm{AB}) \leq H(\mathrm{A}) + H(\mathrm{B})$*, with equality holding iff the joint state of the bipartite system is factorized, namely* $\Pi_{ij} = p_i q_j$*;*

3. *for any* A *and* B *and any* $\Pi \in \mathsf{St}_1(\mathrm{AB})$ *one has* $H(\mathrm{A} : \mathrm{B}) \geq 0$ *with equality holding iff the joint state of the bipartite system is factorized;*

4. *for any* A *and* B *and any* $\Pi \in \mathsf{St}_1(\mathrm{AB})$ *one has*

$$H(\mathrm{A} : \mathrm{B}) \leq \min\{H(\mathrm{A}), H(\mathrm{B})\}; \tag{2.8}$$

5. *(concavity) if* $\mathbf{p}, \mathbf{q} \in \mathsf{St}_1(\mathrm{A})$ *have the same outcome set and* $s \in [0, 1]$ *then*

$$H(s\mathbf{p} + (1-s)\mathbf{q}) \geq sH(\mathbf{p}) + (1-s)H(\mathbf{q}); \tag{2.9}$$

6. *(strong subadditivity) For any tripartite classical system* ABC*, the following inequality holds*

$$H(\mathrm{ABC}) + H(\mathrm{C}) \leq H(\mathrm{AB}) + H(\mathrm{BC}). \tag{2.10}$$

Some of these properties, such as subadditivity and concavity, are desirable features of a quantity that aims at measuring how much information is carried by a physical system.

2.1.3 Weak tipicality and the first Shannon theorem

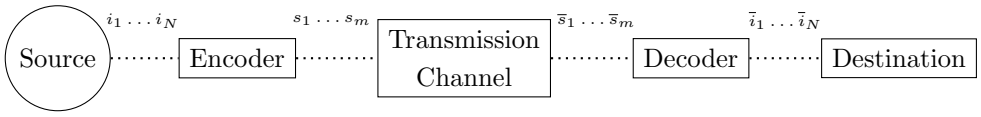


Figure 2.2: A schematic representation of a communication setting. A discrete source generates a sequence of N symbols that are the input of the encoder, used by the sender to generate a string of bits that represents a compressed version of the original message. Then a transmission channel is used to convey the latter to the receiver, that finally decodes the message.

A communication scheme is made of three components: the *information source* that produces the messages received by a first agent, say Alice; the *encoder*, with which the sender compresses the messages on a smaller physical support; the *channel* that is used to transmit the compressed message to a receiver, and which may introduce some noise and distortion of the compressed message (but in the present case we consider noiseless coding); finally, the *decoder* that decompresses the original message for a second agent, say Bob. Thus, the first question is how to represent such basic elements of a communication setting. In particular, one has to model the output of an

information source, and also characterise the result of an operation that is performed thereon. Now, an information source might give rise to signals of any sort, that, for instance, can be represented as a continuum in the case of the sound of a vinyl record produced by turntable speakers, or as a message with symbols drawn from an alphabet of finite size (e.g. a telegraph).

In this section, in the spirit of keeping the discussion as simple as possible, we will be concerned with the easiest case to handle. Imagine that we have a classical source of information whose signals are in the form of sequences $\mathbf{i} = i_1 \dots i_N$ of variable lentgh N, where each symbol i in the sequence is drawn from a given set X, according to a probability distribution $\mathbf{p} := \{p_i\}_{i\in\mathsf{X}}$. If we assume that they are *identically and independently distributed* (i.i.d.), the probability associated with a given string is the product of the individual probabilities $p_{i_1} \cdots p_{i_N} =: p_{\mathbf{i}}$. With our notation, this means that the state that represents a *possible* message is associated with a classical system A^N represented by $(\mathbb{R}^{d_\mathrm{A}})^{\otimes N}$ and it has the form $\rho = \sum_{\mathbf{i}} p_{\mathbf{i}}\ \mathbf{e}_{\mathbf{i}}$, where $\mathbf{e}_{\mathbf{i}} = \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_N}$ are the pure state vectors of the composite system $(\mathbb{R}^{d_\mathrm{A}})^{\otimes N}$.

Remark 2.1.2. [25] A proper way of mathematically modeling the output of a source is as a stochastic process, namely as a sequence of random variables $I_1 \dots I_N, \dots$ with an associated probability distribution $P^{(N)}(I_1 \dots I_N = i_1 \dots i_N)$ for any N, and assuming that the source is discrete means that the symbols i_k are drawn from a discrete set X, usually called *alphabet*. A reasonable assumption is that the probability distributions P^N satisfies the compatibility condition $\sum_{i\in\mathsf{X}} P^{(N+1)}(i_1 \dots i_N i) = P^{(N)}(i_1 \dots i_N)$, in this case one can safely omit the apex N with no risk of confusion and write $P(i_1, \dots, i_N)$ whatever N is. It is then clear that there is a huge variety of options that one might consider. For instance, one can model a source as a Markov chain, where the probability of having the symbol i_N after N steps, given that the preceeding symbols are $i_1 \dots i_{N-1}$, is actually dependent on the last symbol i_{N-1} only, or as a generalizion of this in which conditioned probabilities depend on the last two, or three, or k symbols. In order to avoid a huge amount of technicalities that may cloud the concepts, as we have already said, we will make the simplest assumption on the nature of the source, namely that the random variables $I_1, \dots, I_N, \dots$ are identically and indipendently distributed; in this case, the source is said to be stationary and memoryless.

Once the messages are emitted we can adopt different strategies in order to digitise them. A possibility is to use a variable-length code, where each symbol is coded on strings of bits of different lentgh, via an injective function (so that all the symbols can be recovered without errors). According to the occurence probability of the symbols, the most likely can be encoded onto just one bit, while for the unlikely ones longer bit strings can be chosen. As a simple example, let us consider a source that emits sequences made of symbols drawn from the set $\mathsf{X} = \{a, b, c, d\}$, with probability

$$p(a) = \frac{1}{2}, \quad p(b) = \frac{1}{4}, \quad p(c) = p(d) = \frac{1}{8}.$$

One can encode the symbols according to a bijective function that uses two bits for each symbol, so that the expected number of bits per symbol is, indeed, two. But if we want to exploit the information on the occurrence probability of the symbols, we can use the following injective encoding

$$a \to 1, \quad b \to 10, \quad c \to 110, \quad d \to 111.$$

In this case the average number of bits per symbol is $7/4$, so that this encoding scheme has a better compression rate.

Shannon considered two differences with respect to the simple scenario that we have described so far. In the first place, instead of considering symbol coding, Shannon adopted a *block-coding* strategy. In this case, we wait until the source has emitted a message made of a large number of symbols. Then, we encode the entire strings, instead of the symbols they are made of. The idea behind this approach is that, for fixed N, there is a subset of all the sequences that, while having cardinality which is exponentially smaller than the set of all sequences, it retains almost all the probability. These sets are called "typical" and they are formally defined as follows.

Definition 2.1.2 (Typical set and typical sequences)**.** *For any $\delta > 0$, the (N, δ)-typical set $T_\delta^N(\mathbf{p})$ is a subset of all the possible strings of lentgh N and it is defined as follows*

$$T_\delta^N(\mathbf{p}) = \left\{ \mathbf{i} : \left| \frac{1}{N} \log_2 \frac{1}{p_\mathbf{i}} - H(\mathbf{p}) \right| \leq \delta \right\},$$

where $H(\mathbf{p})$ is the Shannon entropy of $\mathbf{p}$. Accordingly, any $\mathbf{i} \in T_\delta^N(\mathbf{p})$ is called (N, δ)-typical sequence*, or simply* typical sequence *(whenever N and δ are clear from the context).*

The quantity $-1/N \log p_\mathbf{i}$ is also called the *sample entropy* of the string $\mathbf{i} = i_1 \dots i_N$, thus the set $T_\delta^N(\mathbf{p})$ contains those sequences whose sample entropy is close to the Shannon entropy within δ. It is worth mentioning a simple rephrasing of this definition, which is called *equipartition property* of the typical sequences, namely that their probability is almost uniform in the following sense

$$2^{-[NH(\mathbf{p})+\delta]} \leq p_\mathbf{i} \leq 2^{-[NH(\mathbf{p})-\delta]}. \tag{2.11}$$

The following is a summary of the relevant properties of typical sequences.

Theorem 2.1.2. *Let $\mathbf{p}$ be a probability distribution, and for $\delta > 0$ let $T_\delta^N(\mathbf{p})$ denote the (N, δ)-typical set. Then*

1. *Let $\eta > 0$. Then there exists N_0 such that for any $N \geq N_0$*

$$P(\mathbf{i} \in T_\delta^N(\mathbf{p})) \geq 1 - \eta;$$

2. *Let $\eta > 0$. Then there exists N_0 such that, for any $N \geq N_0$, the cardinality of $T_\delta^N(\mathbf{p})$ is bounded as follows*

$$(1-\eta)2^{N[H(\mathbf{p})-\delta]} \leq |T_\delta^N(\mathbf{p})| \leq 2^{N[H(\mathbf{p})+\delta]}.$$

These two properties formalize in a rigorous way what we have already mentioned: first, for any given $\delta > 0$, the set of typical sequences has all the probability in an asymptotic sense; moreover, the number of typical sequences for N sufficiently large and δ sufficiently small is approximetely equal to $2^{NH(\mathbf{p})}$, that is exponentially smaller than $|\mathsf{X}|^N$. The notion of weak typicality is a powerful tool for proving results in classical information theory, and it will suffice for the purposes of the present work. It is relevant not only for problems in classical information theory, but it turns out to be crucial also in the quantum setting, where a notion of typical subspace has been developed relying on that of typical sequences, as we will discuss in the next section.

Insofar we have discussed the nature of the source, and the form of the message that arrives at the input of the Alice's encoder. According to section 2.1, the mathematical form of a possible encoding channel $\mathscr{E}$ is given by a stochastic matrix $E_{\mathbf{s},\mathbf{i}}$ that contains the information on how the input strings $\mathbf{i} \in T_N$ are mapped onto strings of, say, M bits $\mathbf{s} \in S_M$, and it is fully specified by its action on the pure states $\mathbf{e_i}$. In other words, any pure state $\mathbf{e_i}$ is mapped onto a state of the form $\sum_{\mathbf{s}} E_{\mathbf{s},\mathbf{i}} \mathbf{e_s}$. Once the bit strings are sent to Bob (by means of an ideal channel), the latter tries to recover the original message by applying his decoding channel $\mathscr{D}$, represented by a stochastic matrix $D_{\mathbf{i},\mathbf{s}}$, so that he converts the bit string $\mathbf{s}$ to a sequence $\mathbf{i}$ of symbols in X. Therefore, if the input was the pure state $\mathbf{e_i}$, then this has been distorted by the compression-decompression operation, and Bob now has $\sum_{\mathbf{j},\mathbf{s}} D_{\mathbf{j},\mathbf{s}} E_{\mathbf{s},\mathbf{i}} \mathbf{e_j}$ as output from its decoder. Now, if the mapping is made deterministically, that is to say if each string $\mathbf{i}$ is mapped onto a unique $\mathbf{s}$ through $E_{\mathbf{s},\mathbf{i}}$ and similarly for $D_{\mathbf{i},\mathbf{s}}$ ($E_{\mathbf{s},\mathbf{i}} = 1$ for some $\mathbf{s}$ for any $\mathbf{i}$ and correspondingly for $D_{\mathbf{i},\mathbf{s}}$), the final state is a pure state $\mathbf{e_j}$. If the number of available bits is such that $2^M \geq d^N$, then the encoding-decoding scheme can be chosen in such a way that $\mathbf{j} = \mathbf{i}$ for any choice of $\mathbf{i}$[2].

Clearly, Alice can save each string $\mathbf{i}$ on a different bit string $\mathbf{s}$ only if the number of bits M that she has at her disposal is sufficiently high, whence Bob can recover correctly any message. Here comes the second idea of Shannon, that is to use less bits but allow for an error in the decoding. Such an error is quantified by the error probability, namely the probability that the final string is different from the original one. Letting $C_{\mathbf{j},\mathbf{i}} := \sum_s D_{\mathbf{j},\mathbf{s}} E_{\mathbf{s},\mathbf{i}}$ be the transition matrix (representing the classical channel $\mathscr{C} := \mathscr{D}\mathscr{E}$), if the state of the source is $\rho = \sum_i p_i \mathbf{e}_i$, each string is produced with probability $p_{\mathbf{i}}$, and the probability of having the string $\mathbf{j}$ after the encoding-decoding operation is $C_{\mathbf{j}|\mathbf{i}} p_{\mathbf{i}}$. The error

[2]In other words, in this case the encoding and decoding functions can be described by an injective function $h : T_N \to S_M$ and a function that inverts the action of h on $h(T_N)$.

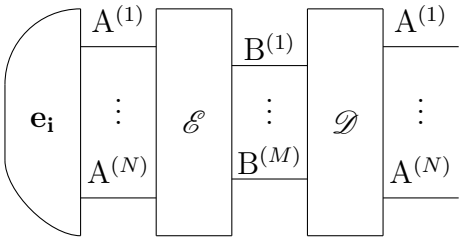


Figure 2.3: A diagrammatic representation of the communication setting. For fixed N, each sequence is represented by the pure state $\mathbf{e_i} \in (\mathbb{R}^{d_\mathrm{A}})^{\otimes N}$. After the coding-decoding operation, the resulting string can be different from the original one, according to the probability distribution $\sum_{\mathbf{s}} D_{\mathbf{j},\mathbf{s}} E_{\mathbf{s},\mathbf{i}}$. The final state on A^N is represented by $\sum_{\mathbf{j},\mathbf{s}} D_{\mathbf{j},\mathbf{s}} E_{\mathbf{s},\mathbf{i}} \mathbf{e_j}$.

probability $p(\mathscr{C}, \rho^{\otimes N})$ is then computed according to the following formula

$$p(\mathscr{C}, \rho^{\otimes N}) := \sum_{\mathbf{i}} \sum_{\mathbf{j}\neq\mathbf{i}} p(\mathbf{j}|\mathbf{i}) p_{\mathbf{i}} = \sum_{\mathbf{i}} \sum_{\mathbf{j}\neq\mathbf{i}} C_{\mathbf{j}|\mathbf{i}} p_{\mathbf{i}}. \tag{2.12}$$

Remark 2.1.3. Notice that the error probability can also be written in terms of the distance in equation (2.2) as follows

$$p(\mathscr{C}, \rho^{\otimes N}) = \frac{1}{2} \sum_{\mathbf{i}} p_{\mathbf{i}} \|\mathscr{C}(\mathbf{e_i}) - \mathbf{e_i}\|_1, \tag{2.13}$$

where $\mathscr{C}(\mathbf{e_i}) = \sum_{\mathbf{j}} C_{\mathbf{j}|\mathbf{i}} \mathbf{e_j}$. This is a straightforward computation that follows by the normalization condition $\sum_{\mathbf{j}} C_{\mathbf{j}|\mathbf{i}} = 1$, that holds for any $\mathbf{i}$. Indeed, first observe that the normalization condition implies

$$p(\mathscr{C}, \rho^{\otimes N}) := \sum_{\mathbf{i}} \sum_{\mathbf{j}\neq\mathbf{i}} p(\mathbf{j}|\mathbf{i}) p_{\mathbf{i}} = 1 - \sum_{\mathbf{i}} C_{\mathbf{i},\mathbf{i}} p_{\mathbf{i}}. \tag{2.14}$$

As we have recalled at the beginning of this section, the set of states of CT is given by a simplex, and any probability vector representing a state can be uniquely decomposed in terms of pure states $\mathbf{e}_i$, corresponding to vectors with all zero components except the one at the i-th position: $(e_i)_j = \delta_{i,j}$. Again, the normalization condition on the columns of C implies the following chain

of equalities

$$\begin{aligned}
&\sum_{\mathbf{i}} p_{\mathbf{i}} \|C\mathbf{e}_{\mathbf{i}} - \mathbf{e}_{\mathbf{i}}\|_1 \\
&= \sum_{\mathbf{i}} p_{\mathbf{i}} \sum_{\mathbf{j}} |C_{\mathbf{j},\mathbf{i}} - \delta_{\mathbf{ij}}| \\
&= \sum_{\mathbf{i}} p_{\mathbf{i}} (\sum_{\mathbf{j}\neq\mathbf{i}} C_{\mathbf{j},\mathbf{i}} + 1 - C_{\mathbf{i},\mathbf{i}}) \\
&= \sum_{\mathbf{i}} 2p_{\mathbf{i}}(1 - C_{\mathbf{i},\mathbf{i}}) = \\
&= 2\sum_{\mathbf{i}} p_{\mathbf{i}}(1 - C_{\mathbf{i},\mathbf{i}}) = 2(1 - \sum_{\mathbf{i}} p_{\mathbf{i}} C_{\mathbf{i},\mathbf{i}}) = 2p(\mathscr{C}, \rho^{\otimes N}),
\end{aligned}$$

namely

$$p(\mathscr{C}, \rho^{\otimes N}) = \sum_{\mathbf{i}} p_{\mathbf{i}} \frac{1}{2} \|C\mathbf{e}_{\mathbf{i}} - \mathbf{e}_{\mathbf{i}}\|_1.$$

Thus, a small error probability corresponds to asking a small average distortion of the state that represents the message emitted by the source.

For fixed N and a given tolerance ε on the error decoding, the efficiency of a given compression scheme is measured by its *rate*, that is given by the ratio M/N between the number of bits M used for the encoding and the length N of the messages. The natural question that now arises is: what is the minimal rate such that a family of compression schemes with asymptotically vanishing error exist in the limit of long messages? Shannon answered to this question, and its result is summarized in the following statement.

Theorem 2.1.3. *Let $\mathbf{p} \in \mathsf{St}_1(\mathrm{A})$ be a classical state describing an i.i.d. (classical) source of information. Then for any $\delta > 0$ and $\varepsilon > 0$ there exists N_0 such that for any $N \geq N_0$ there exists a compression scheme with rate $R \in (H(\mathbf{p}), H(\mathbf{p}) + \delta]$ and $p(\mathscr{C}, \rho^{\otimes N}) < \varepsilon$. Moreover, for any rate $R < H(\mathbf{p})$ and $\varepsilon > 0$ there exists N_0 such that any compression scheme with rate R and $N \geq N_0$ is such that $p(\mathscr{C}, \rho^{\otimes N}) \geq \varepsilon$.*

The first part tells us that we can consider transmission rates arbitrarily close to the Shannon entropy $H(\mathbf{p})$ of the source, and, as long as we are concerned with a sufficiently large length N of the messages, we can achieve an error as small as we wish. This part is usually referred to as the *direct* part, and it establishes the achievability of $H(\mathbf{p})$ as a compression rate. The other part tells us that no compression schemes with rate smaller than $H(\mathbf{p})$ can have arbitrary small error probability, and this establishes the minimality of $H(\mathbf{p})$. The full operational intepretation of the Shannon entropy is guaranteed by the fact that $H(\mathbf{p})$ is both *achievable* and *minimal*.

Remark 2.1.4. Given the notion of weak tipicality introduced in the preceeding subsection, it is clear how to devise a sequence of compression schemes that

works: taking $M = \lceil N[H(\mathbf{p}) + \delta/2] \rceil$, one can consider an encoding channel that deterministically saves the typical strings and maps the others on a fixed string of bits. The decoding is required to precisely invert the bit strings corresponding to the typical one; in this way, the error probability can be made arbitrarily small for sufficiently large N, and the rate M/N is in the required range for N sufficiently high.

2.2 Quantum Shannon theory

While in the previous section we gave an account of the first Shannon theorem, together with all the apparatus needed for its statement, the purpose of this section is to review its quantum analogue. First, we briefly recall the mathematical structure of quantum theory, and the relevant distinguishability measures that are commonly used in quantum information theory, together with the definition of the von Neumann's entropy. We then proceed with a thorough exposition of the quantum noiseless coding theorem, for the case of a source described by a pure state ensemble, also reminding the reader of the main tool necessary for its proof, the notion of typical subspace.

2.2.1 Compendium of Quantum Theory

The description that follows is standard as it is presented in any book on quantum information theory (such as [26, 27]). In quantum theory systems are associated with complex Hilbert spaces, and the type of a system is identified by the dimension d of the associated Hilbert space (we will assume $d < \infty$ whenever we will refer to results relative to quantum theory). Therefore, different types are described by different dimensions, and systems of the same type are equivalent for information-theoretic purposes.

The convex set of states of a system A is mathematically represented by the set of sub-normalized matrices ρ on the associated space $\mathcal{H}_{\mathrm{A}}$, and it is denoted by $\mathsf{St}(\mathrm{A})$. Normalized states correspond to the convex set of density matrices, which is denoted by $\mathsf{St}_1(\mathrm{A})$

$$\begin{aligned} \mathsf{St}(\mathrm{A}) &:= \{\rho \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})_{>} | \operatorname{Tr}(\rho) \leq 1\}, \\ \mathsf{St}_1(\mathrm{A}) &:= \{\rho \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})_{>} | \operatorname{Tr}(\rho) = 1\}. \end{aligned}$$

Pure states are defined as those elements of $\mathsf{St}_1(\mathrm{A})$ that are rank-one, and, accordingly, all the other states are called mixed. The size D of a quantum system A is the dimension of the linear space $\mathsf{St}_{\mathbb{R}}(\mathrm{A})$ spanned by the set of states, therefore it is given by d^2, where d is the dimension of the Hilbert space associated with the system.

For a given quantum system A, the convex set of effects $\mathsf{Eff}(\mathrm{A})$ is given by functionals (acting on the set of states) of the form $\operatorname{Tr}(\cdot E)$, where E is a positive

operator dominated by the identity $0 \leq E \leq I$ and the (unique) deterministic effect is represented by the trace $\mathrm{Tr}(\cdot)$. The observation tests, that in quantum theory are commonly called Positive Operator-Valued Measures (POVM), are represented by collections of operators $\{E_i\}$ satisfying the condition $\sum_i E_i = I$, namely, they must sum to the deterministic effect.

Quantum operations, whose set is denoted by $\mathsf{Transf}(\mathrm{A} \to \mathrm{B})$, are mathematically represented by Completely Positive Trace non-Icreasing maps, namely, if $\mathscr{C} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$, then

1. (trace non-increasing) $\forall \rho \in \mathsf{St}(\mathrm{A}), \qquad \mathrm{Tr}[\mathscr{C}(\rho)] \leq \mathrm{Tr}(\rho)$;

2. (complete positivity) for any ancillary sistem E, $(\mathscr{C} \otimes \mathscr{I}_{\mathrm{E}})(\Gamma) \geq 0$, for any $\Gamma \in \mathsf{St}(\mathrm{AE})$.

When in the first condition the equality is satisfied for any $\rho \in \mathsf{St}(\mathrm{A})$, $\mathscr{C}$ is said to be Trace Preserving (TP); in particular, determinisitc transformations are all and only those maps that are Completely Positive and Trace Preserving (CPTP). For any quantum transformation $\mathscr{C} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ there exists a set of operators $\{C_i\} \subseteq \mathcal{L}(\mathcal{H}_{\mathrm{B}}, \mathcal{H}_{\mathrm{A}})$, called Kraus operators, such that

$$\mathscr{C}(\rho) = \sum_i C_i^\dagger \rho C_i \qquad \forall \rho \in \mathsf{St}(\mathrm{A}) \tag{2.15}$$

and satisfying the condition $\sum_i C_i^\dagger C_i \leq I$, as a consequence of the Trace non-Increasing condition that must hold for $\mathscr{C}$. Accordingly, reversible transformations are all and only CPTP maps with a single Kraus operator that is also unitary, while left reversible transformations are represented by isometric operators, namely $V : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{A}}$ such that $V^\dagger V = I_{\mathcal{H}_{\mathrm{B}}}$.

Quantum operations can be performed sequentially, as long as the output system of the first operation corresponds to the input system of the second one. This operation is mathematically represented by the composition of maps. Quantum systems A and B can also be composed in parallel to obtain a *bipartite* system AB. Then, the tensor product space $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$ is the Hilbert space associated with the new system AB. In the latter case, when one is interested to only one of the subsystems, either the partial trace Tr_{A} or Tr_{B} must be applied in order to discard the corresponding system.

A distinctive trait of quantum theory is that, given any quantum system A, any mixed state $\rho \in \mathsf{St}_1(\mathrm{A})$ can be *purified*. We state this property in the following theorem for future reference.

Theorem 2.2.1 (Existence and essential uniqueness of purification)**.** *For any ρ there exists a reference system* R*, also called the* purifying system*, and a pure entangled state $\Psi \in \mathsf{St}_1(\mathrm{AR})$ such that*

$$\rho = \mathrm{Tr}_{\mathrm{R}}(\Psi). \tag{2.16}$$

Moreover, given two different purifications of the same state ρ, say Ψ and Ψ', there always exists a quantum channel $\mathscr{C} \in \mathsf{Transf}(\mathrm{R} \to \mathrm{R}')$ such that

$$\Psi' = (\mathscr{I}_{\mathrm{A}} \otimes \mathscr{C})(\Psi). \tag{2.17}$$

When R *and* R′ *are of the same dimension, then $\mathscr{C}$ is reversible.*

Equation (2.16) affirms the *existence* of a purification for any A and any state $\rho \in \mathsf{St}(\mathrm{A})$, while equation (2.17), in the case of systems R, R′ of the same type, asserts the *essential uniqueness* of purification up to reversible channel applied to the purifying system.

We conclude by introducing the double-ket notation [28]. Denoting by $\{e_i\}_{i=1}^{d_{\mathrm{A}}}$ and $\{f_j\}_{j=1}^{d_{\mathrm{B}}}$ a pair of orthonormal bases of the Hilbert spaces $\mathcal{H}_{\mathrm{A}}$ and $\mathcal{H}_{\mathrm{B}}$ respectively, we define the linear operator $\mathsf{Vec} : \mathcal{L}(\mathcal{H}_{\mathrm{A}}, \mathcal{H}_{\mathrm{B}}) \to \mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{A}}$ as follows

$$\mathsf{Vec}: \quad A \to |A\rangle\!\rangle := \sum_{i=1}^{d_{\mathrm{A}}} \sum_{j=1}^{d_{\mathrm{B}}} A_{j,i} \, |f_j\rangle \langle e_i| \, . \tag{2.18}$$

This operator is an isomorphism with obvious inverse given by

$$\mathsf{Vec}^{-1}: \quad |\Psi\rangle\!\rangle \to \Psi := \sum_{i=1}^{d_{\mathrm{A}}} \sum_{j=1}^{d_{\mathrm{B}}} \Psi_{j,i} \, |f_j\rangle \langle e_i| \, . \tag{2.19}$$

The following theorem collects some computational properties of the double-ket notation that will be useful in chapter 5

Theorem 2.2.2. *The following identity holds for any $C : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$, $A : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}}$ and $B : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{D}}$.*

$$A \otimes B|C\rangle\!\rangle = |ACB^T\rangle\!\rangle. \tag{2.20}$$

Moreover, for any pair $A, B : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ the following hold

$$\begin{aligned} \mathrm{Tr}_{\mathrm{A}}[|A\rangle\!\rangle\langle\!\langle B|] &= AB^\dagger, \\ \mathrm{Tr}_{\mathrm{B}}[|A\rangle\!\rangle\langle\!\langle B|] &= A^T B^*, \\ \mathrm{Tr}[|A\rangle\!\rangle\langle\!\langle B|] &= \langle\!\langle B|A\rangle\!\rangle = \mathrm{Tr}[B^\dagger A] = \mathrm{Tr}[B^* A^T]. \end{aligned}$$

Remark 2.2.1. The usual mathematical description of QT presented so far can also be derived as a consequence of three Hilbert-space postulates. The first two ones concern the nature of the systems and the way they compose, while the third indirectly tells us what are the transformations of the theory.

Postulate 1. *We associate each quantum system* A *with a complex Hilbert space $\mathcal{H}_{\mathrm{A}}$. The composition of two systems* A *and* B*, denoted by* AB*, is associated with the tensor product $\mathcal{H}_{\mathrm{AB}} = \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$.*

Postulate 2. *Any state $\omega \in \mathsf{St}(\mathrm{A})$ is in a one-to-one correspondence with positive operators $\rho_\omega \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})$ satisfying the condition $\mathrm{Tr}(\rho_\omega) \leq 1$. Such a correspondence preserves convex combination.*

Postulate 3. *All the maps that satisfy all the mathematical requirements for representing a transformation within the theory (that are called admissible maps) will be actual transformations of the theory.*

The third postulate implicitly defines the form of effects and transformations. First of all, one can derive the fact that any effect is represented by an operator $E \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})$ that is positive and dominated by the identity; moreover, *any* such operator gives rise to a probability functional, and postulate 3 then entails that it is an actual effect. Concerning the transformations, a map is admissible if it is linear, completely positive and trace non increasing, and the postulate implies that *any* map satisfying these requirements is a legitimate quantum operation.

2.2.2 Distinguishability measures in quantum theory and the von Neumann entropy

Given two states ρ and σ, we can assess how much these are different by using the so called trace-norm distance (or schatten 1-norm). This is well defined for every linear operator $M \in \mathcal{L}(\mathcal{H}_{\mathrm{A}}, \mathcal{H}_{\mathrm{B}})$ between finite dimensional Hilbert spaces. If we set $|M| = \sqrt{M^\dagger M}$, called the modulus of M, the trace-norm of M is defined as follows

$$\|M\|_1 := \mathrm{Tr}\,|M|, \tag{2.21}$$

and the trace-norm distance is the one induced by this norm. The obtained distance is commonly normalized, so that one defines the normalized trace distance between quantum states as

$$d(\rho, \sigma) := \frac{1}{2}\|\rho - \sigma\|_1. \tag{2.22}$$

This is called normalized in that $d(\rho, \sigma) \leq 1$ for any pair of quantum states.

Remark 2.2.2. The distance defined in equation (2.22) has also the following characterization

$$d(\rho, \sigma) = \max_{0 \leq E \leq I} \mathrm{Tr}[E(\rho - \sigma)]. \tag{2.23}$$

This formula allows for an operational interpretation of the normalized distance, that we now recall. Suppose that an agent, say Bob, prepares one of two quantum states, ρ_0 and ρ_1, with equal probabilites $\frac{1}{2}$. A second agent, say Alice, can perform a measurement $\{E_0, E_1\}$ at her wish and then guess which state has been prepared by Bob. In particular, she guesses that the state prepared by Bob is ρ_0 if the outcome of the measurement is "0", and ρ_1 otherwise. Therefore, given the condition $E_0 + E_1 = I$, the maximum success probability is found to be

$$p_{\text{succ}} = \frac{1}{2}\left[1 + \frac{1}{2}\|\rho_0 - \rho_1\|_1\right]. \tag{2.24}$$

Namely, the greater is the success probability for Alice, the greater is the distance between states, and viceversa.

Another quantity that is commonly used in quantum information theory, in order to establish how much two quantum states are different, is the Uhlmann fidelity. For any pair of density matrices ρ and σ this is defined as follows [29]

$$F(\rho,\sigma) := \operatorname{Tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}} = \|\sqrt{\rho}\sqrt{\sigma}\|_1. \tag{2.25}$$

Notice that for pure states $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$, F measures the overlap between the vector states, $F(\rho,\sigma) = |\langle\psi|\phi\rangle|$. Moreover, if $\rho = |\psi\rangle\langle\psi|$ and σ is a density matrix, one has

$$F(|\psi\rangle\langle\psi|, \sigma) = \langle\psi|\sigma|\psi\rangle^{1/2}. \tag{2.26}$$

From the definition it is clear that F is symmetric in its arguments, and that $F(\rho,\sigma) = 1$ whenever ρ and σ coincide, while $F(\rho,\sigma) = 0$ if and only if they have orthogonal support. Mathematically, this quantity is not a distance, but it is equivalent to $d(\rho,\sigma)$ thanks to the well-known Fuchs-van der Graaf [30] inequalities

$$1 - F(\rho,\sigma) \leq \frac{1}{2}\|\rho,\sigma\|_1 \leq \sqrt{1 - F^2(\rho,\sigma)}. \tag{2.27}$$

The fidelity $F(\rho,\sigma)$ also has an alternative characterization: given two density matrices $\rho,\sigma \in \mathsf{St}_1(\mathrm{A})$ and a POVM $\{E_j\}$, one can define the probability distributions $\{p_j\} := \{\operatorname{Tr}(\rho E_j)\}$ and $\{q_j\} := \{\operatorname{Tr}(\sigma E_j)\}$ and compute the classical fidelity between them. It turns out that a minimization over *all* the possible POVMs leads to the Uhlmann fidelity between ρ and σ

$$F(\rho,\sigma) = \min_{\{E_j\}} \sqrt{\operatorname{Tr}(\rho E_j)\operatorname{Tr}(\sigma E_j)}. \tag{2.28}$$

Remark 2.2.3. The proof of this charactersation of the Uhlmann fidelity can be found in [26]. It is worth mentioning that the minimum is achieved when the elements of the POVM correspond to the eigenvector of the operator

$$M := \rho^{-\frac{1}{2}}\sqrt{\rho^{\frac{1}{2}}\sigma\rho^{\frac{1}{2}}}\rho^{-\frac{1}{2}}. \tag{2.29}$$

This observation turns out to be of great importance when we will deal with FT in chapter 5.

Finally, from the Uhlmann fidelity we can define another quantity that is called *entanglement fidelity*. This Measures how much entanglement between two quantum systems A and B is preserved by a quantum channel when the

latter is applied to A (or B). To be more precise, let ρ be a density matrix on A and $\mathscr{C}$ a quantum channel acting on A. Let Ψ be a purification of ρ and B denote the ancillary system. The entanglement fidelity (formerly introduced by Schumacher [31]) is then defined in the following way

$$F(\rho, \mathscr{C}) := \langle\langle \Psi | (\mathscr{C} \otimes \mathscr{I})(\Psi) | \Psi \rangle\rangle. \tag{2.30}$$

This quantity is well defined since it is independent of the particular chosen purification. Now, let $\{C_i\}$ be the kraus operators of the channel $\mathscr{C}$, then $F(\rho, \mathscr{C})$ can also be rewritten in terms of the operators $\{C_i\}$ as follows

$$F(\rho, \mathscr{C}) = \sum_i |\operatorname{Tr}(\rho C_i)|^2. \tag{2.31}$$

This is a useful representation that will be used in this work, specifically when we will deal with FT. We also notice that the entanglement fidelity can be computed as the square of the Uhlmann one between the purification Ψ and the state obtained after the application of $\mathscr{C}$, $\mathscr{C} \otimes \mathscr{I}(\Psi)$

$$F(\rho, \mathscr{C}) = F^2[|\Psi\rangle\langle\langle\Psi|, (\mathscr{C} \otimes \mathscr{I})(|\Psi\rangle\langle\langle\Psi|)]. \tag{2.32}$$

The quantum notion of entropy has been formerly introduced by John von Neumann in the first german edition of [22], as a generalization of the Gibbs entropy of statistical mechanics. It also generalizes to the quantum scenario the Shannon entropy (despite its earlier introduction). We recall here its mathematical definition.

Definition 2.2.1. *Given a state on a quantum system* A *represented by a density matrix ρ, the von Neumann entropy of ρ is defined as:*

$$S(\rho) := -\operatorname{Tr}(\rho \log_2 \rho). \tag{2.33}$$

Whenever the state is clear from the context, we will simply write $S(\mathrm{A})$.

When we consider the spectral decomposition of ρ, $S(\rho)$ reduces to the Shannon entropy of the probability distribution defined by its eigenvalues. More precisely, if $\rho = \sum_i p_i |x_i\rangle \langle x_i|$, where p_i are the eigenvalues of ρ (satisfying $\sum_i p_i = 1$) and $|x_i\rangle$ are the corresponding eigenstates, then it holds $S(\rho) = -\sum_i p_i \log_2 p_i$. The definition also applies when bipartite systems AB are considered: if Π is a density matrix representing a state of AB, its von Neumann entropy is simply given by $S(\Pi) = -\operatorname{Tr}(\Pi \log_2 \Pi)$, and the marginal entropies are obtained as the von Neumann entropies $S(\rho_{\mathrm{A}})$ and $S(\rho_{\mathrm{B}})$ of the marginal states $\rho_{\mathrm{A}} = \operatorname{Tr}_{\mathrm{B}} \Pi$ and $\rho_{\mathrm{B}} = \operatorname{Tr}_{\mathrm{A}} \Pi$ respectively. The Quantum mutual entropy is defined as $S(\mathrm{A} : \mathrm{B}) = S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{AB})$, analogously to the classical case (see formula (2.7)). The von Neumann entropy and quantum mutual entropy satisfy all the properties listed in theorem 2.1.1 for the Shannon entropy.

2.2.3 Quantum weak typicality and the noiseless coding theorem

A communication scheme, as we have already described before, is essentially given by four elements: an information source, an encoder, a decoder, and the quantum channel that is used in order to transmit the compressed messages, received by the source, from the encoder to the decoder. The latter is assumed to be noiseless, whence it is mathematically described by the identity map, and the encoding and decoding operations are the only source of noise. In [21] the quantum source is modeled as an ensemble of pure states $\{p_i, |\phi_i\rangle \langle\phi_i|\}_{i\in\mathsf{X}}$, not necessarily orthogonal, in such a way that the message for a single usage is given by the density matrix $\rho = \sum_{j\in\mathsf{Y}} p_j |\phi_j\rangle \langle\phi_j|$. Therefore, an information source is essentially defined by a quantum system A and a density matrix $\rho \in \mathsf{St}_1(\mathrm{A})$ on that quantum system. On the other hand, a mixed state also arises when we are in the presence of a bipartite quantum system that is in entangled pure state. This second point of view, already considered in [21], is particularly enlightening, since it makes manifest that *transposing quantum information faithfully means preserving quantum entanglement* between the source and a reference.

The aim of a compression protocol is to use a limited number of resources in order to store information from a quantum system and transmit it from a sender to a receiver. To this end one can encode the message ρ, by means of a quantum operation $\mathscr{E}$, on another system, say B, that has a dimension that is smaller than the orginal one. But, unless the quantum operation is reversible, the original message is unlikely to be recovered in a perfect way, once the decoding operation $\mathscr{D}$ is applied to the compressed message $\mathscr{E}(\rho)$. Since what we are actually interested in is preserving the entanglement, the natural measure for assessing the goodness of the scheme, defined by the pair $(\mathscr{E}, \mathscr{D})$, is the entanglement fidelity. If we let d be the dimension of the Hilbert space $\mathcal{H}_\mathrm{B}$ associated with system B, the following scenarios may occur:

a. if the message ρ has a small overlap with all projectors P onto d-dimensional subspaces of $\mathcal{H}_\mathrm{A}$, that is to say $\mathrm{Tr}(P\rho) < \varepsilon$, then the fidelity is bounded from above, i.e. $F(\rho, \mathscr{D}\mathscr{E}) < \varepsilon$;

b. if there exists at least one projector P onto a d-dimensional subspace that has a high overlap with the message ρ, namely $\mathrm{Tr}(P\rho) > 1 - \varepsilon$, then the fidelity can be made correspondingly high, precisely $F(\rho, \mathscr{D}\mathscr{E}) > 1 - 2\varepsilon$.

The first fact follows by the Cauchy-Schwarz inequality and formula (2.31) for the entanglement fidelity. The second one is proven by using the simple protocol proposed by Schumacher. First, we perform a two outcome test, where we project onto the subspace $\mathcal{K}$ defined by P, corresponding to the outcome "0", or onto the orthogonal complement $\mathcal{K}^\perp$, corresponding to outcome "1". Conditioned on this test, we apply a partial isometry V that embeds in $\mathcal{H}_\mathrm{B}$ the projection of ρ onto $\mathcal{K}$ (there is enough room in $\mathcal{H}_\mathrm{B}$ for this) if "0" occurs,

otherwise, if the outcome is "1", we a apply measure and prepare channel with a fixed target state $\sigma \in \mathsf{St}_1(\mathrm{B})$. The encoding $\mathscr{E}$ can then be explicitly written as follows

$$\mathscr{E}(\rho) = VP\rho PV^\dagger + \mathrm{Tr}[(I-P)\rho]\sigma, \tag{2.34}$$

while the decoding simply inverts the isometry V

$$\mathscr{D}(\tau) = V^\dagger \tau V. \tag{2.35}$$

This yields the desired bound on $F(\rho, \mathscr{D}\mathscr{E})$, provided that $\mathrm{Tr}(\rho P) > 1-\varepsilon$.

Remark 2.2.4. These statements can be rephrased in terms of the eigenvalues of the message ρ. Indeed, if we denote by $\{|x_i\rangle\}_{i\in\mathsf{X}}$ the eigenvectors of ρ and $\{p_i\}_{i\in\mathsf{X}}$ the corresponding eigenvalues, then one has

$$\mathrm{Tr}(\rho P) = \sum_{i\in\mathsf{X}} p_i \,\mathrm{Tr}[|x_i\rangle\langle x_i| P] = \sum_{i\in\mathsf{X}} p_i \langle x_i| P |x_i\rangle ,$$

that is maximized when P projects onto the subspace spanned by those eigenvectors corresponding to the d greatest eigenvalues $\{\lambda_i\}_{i\in\mathsf{G}}$. Let Λ denote the projector onto such a subspace. Then, one has

$$\mathrm{Tr}(P\rho) \leq \mathrm{Tr}(\Lambda\rho) = \sum_{i\in\mathsf{G}} \lambda_i$$

for any projector P of rank d. Therefore, consider

- **a***. if the message ρ is such that $\sum_{i\in\mathsf{G}} p_i < \varepsilon$, then the fidelity is bounded from above, i.e. $F(\rho, \mathscr{D}\mathscr{E}) < \varepsilon$;
- **b***. if the the message ρ is such that $\sum_{i\in\mathsf{G}} p_i > 1-\varepsilon$, then the fidelity can be made correspondingly high, precisely $F(\rho, \mathscr{D}\mathscr{E}) > 1-2\varepsilon$.

a* and **b*** are conditions equivalent to **a** and **b**.

As in the classical setting, also in the quantum case block-coding is considered. A message of lentgh N consists of an integer number N of letters, each extracted from the same ensemble of pure states, i.e. it is represented by a factorized state of the form $\rho^{\otimes N} := \rho \otimes \cdots \otimes \rho$ on the extended system A^N (associated with the Hilbert space $\mathcal{H}_{\mathrm{A}}^{\otimes N}$). While classical information is measured in terms of bits, quantum information is quantified through its analogue, the qubit. Accordingly, a compression scheme is given by a pair of quantum channels $\mathscr{E} : \mathrm{A}^N \to \mathrm{Q}^M$ and $\mathscr{D} : \mathrm{Q}^M \to \mathrm{A}^N$, where M is a suitable integer number M of qubits intended to store the message. The task is to preserve the quantum entanglement between the quantum system A^N and some reference system R, and the efficiency in accomplishing this goal is measured by the *rate*, M/N.

The problem addressed by Schumacher was to find the smallest achievable rate that allows for a faithful transposition of the entanglement, at least

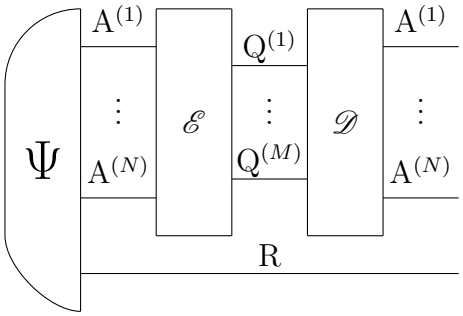


Figure 2.4: A diagrammatic representation of the communication scheme in the quantum setting, as described in the main text. The message $\rho^{\otimes N}$, that lies on the system A^N, to which we have local access, is the marginal of an entangled pure state that purifies it, with R being the purifying system. We encode the system A^N on a number M of qubits with a quantum channel $\mathscr{E}$, then they are transmitted to the receiver, who finally applies a decoding $\mathscr{D}$. The goal is to have a high entanglement fidelity.

asymptotically. As discussed in remark 2.2.4, the problem is connected to the eigenvalues of the messages. The idea of typical subspace considered by Schumacher can be summarised as follows: First, notice that the eigenvalues of $\rho^{\otimes N}$ are of the form $p_{i_1} \dots p_{i_N}$, where $i_k \in \mathsf{X}$ for any $k \in \{1, \dots, N\}$, and p_i can be regarded as a probability distribution associated with a classical random variable with X as outcome set; the corresponding eigenvectors are $|x_{\mathbf{i}}\rangle := |x_{i_1}\rangle \otimes \cdots \otimes |x_{i_N}\rangle$, with $\mathbf{i} := i_1 \cdots i_N$. Therefore, formally, the situation is analogous to what we had in the classical scenario, where $p_{i_1} \dots p_{i_N}$ can be seen as the probability associated with the sequence $i_1 \dots i_N$. It is then natural to distinguish between "typical eigenvectors" and "atypical eigenvectors" whenever the corresponding sequence is either typical or not, respectively; the former set of vectors will then span a subspace that has a dimension approximately equal to $2^{NS(\rho)}$ (since $S(\rho) = -\sum_{i \in \mathsf{X}} p_i \log_2 p_i$) and such that the sum of the eigenvalues is close to 1. The formal definition of typical subspace then follows.

Definition 2.2.2. *Let $\rho \in \mathcal{L}(\mathcal{H})$ be a quantum state given by a density matrix. The* (N, δ)-typical subspace *of ρ is the subspace spanned by the following set of vectors in $\mathcal{H}^{\otimes N}$*

$$T_\delta^N(\rho) := \left\{ |x_{\mathbf{i}}\rangle : \left| \frac{1}{N} \log_2 \frac{1}{p_{\mathbf{i}}} - S(\rho) \right| < \delta \right\}. \tag{2.36}$$

*Accordingly, any vector $|x_{\mathbf{i}}\rangle \in T_\delta^N(\rho)$ will be called a (N, δ)-*typical vector*, and the typical subspace is given by*

$$\mathsf{H}_\delta^N(\rho) := \mathsf{Span}(T_\delta^N(\rho)). \tag{2.37}$$

In other words, a vector $|x_{\mathbf{i}}\rangle$ is (N,δ)-typical iff the string $\mathbf{i} \in T_\delta^N(\mathbf{p})$, given that $S(\rho) = H(\mathbf{p})$. The projector on the (N,δ)-typical subspace, that we accordingly call the (N,δ)-*typical projector*, is simply given by

$$P_N^\delta(\rho) := \sum_{\mathbf{i}\in T_\delta^N(\mathbf{p})} |x_{\mathbf{i}}\rangle \langle x_{\mathbf{i}}| . \tag{2.38}$$

The overlap between the message $\rho^{\otimes N}$ and the (N,δ)-typical projector is given by the sum of the eigenvalues associated with the typical sequences in $T_\delta^N(\mathbf{p})$, according to the following formula

$$\mathrm{Tr}[P_N^\delta(\rho)\rho^{\otimes N}] = \sum_{\mathbf{i}\in T_\delta^N(\mathbf{p})} p_{\mathbf{i}}.$$

As we have already observed, the typical subspaces have almost all the overlap with $\rho^{\otimes N}$ for sufficiently high N, and their dimension is roughly given by $2^{NS(\rho)}$. Moreover, when the dimension of a subspace is strictly smaller than $2^{NS(\rho)}$, the overlap of $\rho^{\otimes N}$ can be made small for large N. All these observations are collected in the following theorem (see [26] or [27] for the proof).

Theorem 2.2.3. *Let $\rho \in \mathcal{L}(\mathcal{H})$ be a quantum state described by a density matrix. Then the following statements hold:*

1. *For every $\epsilon > 0$ and $\delta > 0$ there exists N_0 such that for every $N \geq N_0$*

$$\mathrm{Tr}[P_N^\delta(\rho)\rho^{\otimes N}] \geq 1 - \epsilon. \tag{2.39}$$

2. *For every $\epsilon > 0$ and $\delta > 0$ there exists N_0 such that for any $N \geq N_0$ the dimension of the typical subspace H_δ^N is bounded as follows*

$$(1-\epsilon)2^{N[S(\rho)-\delta]} \leq \dim[H_\delta^N(\rho)] \leq 2^{N[S(\rho)+\delta]}, \tag{2.40}$$

3. *For given N, let S_N denote an arbitrary orthogonal projection on a subspace of $\mathcal{H}^{\otimes N}$ with dimension $\mathrm{Tr}(S_N) < 2^{NR}$, and with $R < S(\rho)$ fixed. Then for every $\epsilon > 0$ there exists N_0 such that for every $N \geq N_0$ and every choice of S_N*

$$\mathrm{Tr}[S_N\rho^{\otimes N}] \leq \epsilon. \tag{2.41}$$

We have set up all the terminology that is needed for stating the quantum version of the noiseless coding theorem, and we have recalled all the stuffs useful for its proof.

Theorem 2.2.4. *Let ρ be a density matrix describing an i.i.d. quantum information source. Then, for any $\delta > 0$ and $\varepsilon > 0$ there exists N_0 such that for any $N \geq N_0$ there exists a compression scheme with rate $R \in (S(\rho), S(\rho) + \delta]$*

and $F(\rho^{\otimes N}, \mathscr{C}) > 1 - \varepsilon$*. Moreover, given any rate* $R < S(\rho)$ *and* $\varepsilon > 0$ *there exists* N_0 *such that any compression scheme with rate* R *and* $N \geq N_0$ *is such that* $F(\rho^{\otimes N}, \mathscr{C}) \leq \varepsilon$*.*

The statement, *mutatis mutandis*, is exactly the same as its classical analogous; the quantum entropy $S(\rho)$ replaces the Shannon one, and the entanglement fidelity is used in place of the error probability for establishing the goodness of compression protocols. The direct part, namely that $S(\rho)$ is an achievable compression rate, is proven by exhibiting explicitly a family of compression schemes, whose construction is based on the quantum notion of weak typicality.

Remark 2.2.5. Since the quantum protocol for achievability serves as a basis for the fermionic case, it is worth describing it in a little more detail, and it has essentially the form given by equations (2.34) and (2.35), where we use the projector onto the typical subspace. Given $\delta > 0$, for fixed N one can take a number of qubits $M = \lceil N[S(\rho) + \delta/2] \rceil$, and for N sufficiently large M/N falls in the desired interval. With this number of qubits, there exists a family of compression schemes that, for any $\varepsilon > 0$ and sufficiently large N, are such that $\mathrm{Tr}[P_N^\delta(\rho)\rho^{\otimes N}] > 1 - \varepsilon$. The theorem on typical subspaces immediately suggests the working strategy: first, by item 1 we deduce that a measurement that asks whether the message has support on the typical subspace yields an answer that is almost "yes" as N becomes large. Given the choice of M and the bound in 2 on $\dim[H_{\delta/2}^N(\rho)]$, there is enough room in order to store all the typical subspace on M qubits. Therefore, a successful protocol is obtained by encoding all the support of $\rho^{\otimes N}$ that lives in the typical subspace and throwing away all the rest. In other words, first a test is performed where $\rho^{\otimes N}$ is projected either onto the typical subspace, via the typical projector $P_N^{\delta/2}(\rho)$, or onto its orthogonal complement, by means of $I - P_N^{\delta/2}(\rho)$. Then, if the outcome of the test corresponds to the projection onto $H_{\delta/2}^N(\rho)$, we isometrically embed the projected state onto M qubits, otherwise we discard and we prepare some fixed state in the typical subspace. The decoder $\mathscr{D}$ then simply inverts the isometry. Formally, the quantum channel corresponding to the encoding acts as

$$\mathscr{E}(\sigma) := V P_N^\delta(\rho)\sigma P_N^\delta(\rho) V^\dagger + \mathrm{Tr}[(I - P_N^\delta(\rho))\sigma] \left|x_{\mathbf{i}}\right\rangle \left\langle x_{\mathbf{i}}\right|, \quad \forall \sigma \tag{2.42}$$

where $V : \mathcal{H}_{\mathrm{A}}^{\otimes N} \to \mathcal{H}_{\mathrm{Q}}^{\otimes M}$ is the partial isometry with support on $H_{\delta/2}^N(\rho)$, namely

$$V^\dagger V = P_N^{\delta/2}(\rho). \tag{2.43}$$

The decoding is simply given by $\mathscr{D}(\tau) := V \tau V^\dagger$. With this protocol, 1 guarantees that the entanglement fidelity will be definitely close to 1, and the compression schemes have the desired rate. The converse part is proven by considering

the most general encoding-decoding pair $(\mathscr{E}, \mathscr{D})$ allowed by quantum theory, and using item 3, along with the fact that $F(\rho^{\otimes N}, \mathscr{D}\mathscr{E}) \leq \mathrm{Tr}(S_N \rho^{\otimes N})$[3].

Remark 2.2.6. Here we have exposed the problem of quantum coding for the transmission of ensembles of pure states, that historically was the first to be solved in the literature. The proof in the original reference [21] for the converse part was given considering unitary decodings only. In [24], the authors pointed out that, by using non unitary decodings, one can obtain a better fidelity, but they proved that, even if we allow for general decoding channels, $S(\rho)$ is the optimal compression rate. For further refinements on results concerning the case of pure state ensembles, see also the Ph.D. thesis of Andreas Winter [32] (in particular, chapter 1). The case of quantum sources modeled as an ensemble of mixed states was considered in [33, 34, 35] and finally solved in [36]. Further generalizations can be considered, as those described, for instance, in the more recent works [37, 38].

[3]This is a consequence of the Cauchy-Schwarz inequality applied with formula (2.31)

CHAPTER 3

The Operational Probabilistic Theories framework

Over the last two decades, the operational approach to foundations of physics has been revitalised by a deeper comprehension of QT, mostly due to theoretical results that have been discovered in the field of quantum information theory. This knowledge have been used in the Generalised Probabilistic Theories (GPT) [39, 40, 41, 42] and OPT [43, 44] frameworks, where various authors have studied the borders of QT from a more general perspective. In the first place, several axiomatisations of QT have been proposed, where the Hilbert space formulation has been derived by a set of information-theoretic axioms[1] [16, 17]. Moreover, several features that characterises quantum information theory have been studied in such frameworks, and it has been proven that they actually show up in more general scenarios, that are not strictly quantum. Notable examples are no-cloning [50], quantum teleportation [51] and the tradeoff between information and disturbance [52], just to mention a few. OPTs also allow for the study of conceptual facets of physics. For instance, in [53] the authors exhibited a probabilistic toy-theory that disentangles the notion of determinism from that of causality, while in [44] it is shown that the presence of entanglement does not necessarily imply complementarity in a physical theory.

In this chapter we summarise such a framework, starting from a description of the operational and probabilistic language. We recall a set of axioms [17]

[1] See also [45, 46] and [47, 48, 49] for different mathematical approaches.

that pick out QT as a special instance of the framework, and, in particular, we introduce two features that are relevant in our treatement, in particular in chapter 4 with consequences for chapters 5 and 6 also: *conditioning* (or *strong causality*), and *steering*. We then provide a brief recap of fermionic information theory, that will be treated in chapter 5, and Bilocal Classical Theory, considered in chapter 6. We finally review the definition and the main properties of the three different definitions of entropy already introduced in the literature [54, 55, 56], along with a generalisation of fidelity [54].

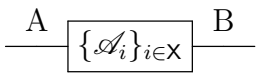
3.1 General description

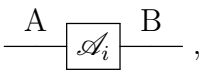
The mathematical framework of OPTs can be conceived as an abstraction of the basic elements and operations of an experimental setting. A complete experimental apparatus consists of a *preparation* of a physical system in some state, a sequence of *operations* performed thereon, and finally making an *observation*. Preparations, operations and observations, are therefore the *events* of a physical experiment, each extracted from a *test* and connected by *systems*. These are the primitive entities of an operational theory. Then, defining the latter, means declaring what are the tests, whence the events, and the systems, and giving a prescription for composing systems, in sequence or in parallel. Probabilities finally come into play, since we are dealing with scientific experiments, and what we are actually interested in is making predictions on what may happen when we have a complete experiment, that starts with a preparation and ends with an observation.

In this section we briefly remind the reader of the basic structures of OPTs, following the presentation of [43, 57]. A detailed account that partially links the framework with category theory can be found in [44].

3.1.1 Primitive notions and the operational language

The primitive notions of an operational theory are those of test, event and system. A *test* $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ is given by a collection of *events*, where i labels the elements of the *outcome space* X (which will be assumed to be finite), that contains the outcomes associated with the test. The *systems* allow for the connection between different tests, and are denoted by capital Roman letters $\mathrm{A}, \mathrm{B}, \ldots$. Therefore, a test is completely determined by its input and output systems, and the events associated with the outcome space X. In order to represent a test and its events $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ we use the usual diagrammatic notation

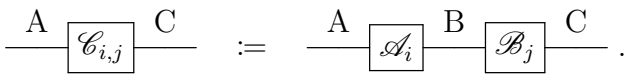


,


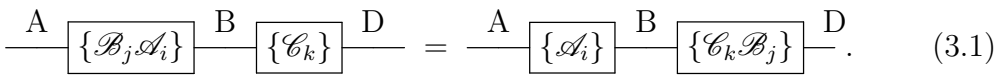
,

and we will call A, B the input and the output system of the test, respectively. Accordingly, the set of events from the input system A to the output system B

will be denoted by $\mathsf{Event}(\mathrm{A} \to \mathrm{B})$. If $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ and $\{\mathscr{B}_j\}_{j\in\mathsf{Y}}$ are two tests, one can define their *sequential composition* as the test $\{\mathscr{C}_{i,j}\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$, with events $\mathscr{C}_{i,j} \equiv \mathscr{B}_j\mathscr{A}_i$ that are diagrammaticaly represented by

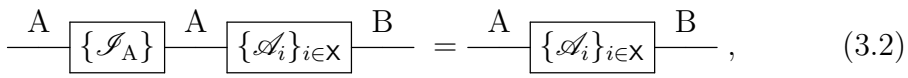


Notice that this definition requires the output system of the events on the left to be necessarily the input system of the events on the right. Since the order in which different operations are sequentially composed does not correspond to an actual physical process, this operation is required to be associative, namely, for all tests $\{\mathscr{A}_i\}$, $\{\mathscr{B}_j\}$, $\{\mathscr{C}_k\}$ and systems A, B, C, D, the following holds

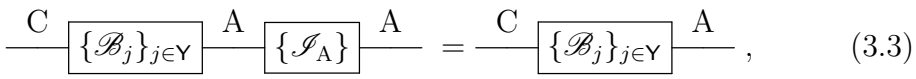


(3.1)

A *singleton test* is a test whose outcome space set X is a singleton, and the unique event contained in it is called *deterministic*. For any system A there exists a unique deterministic *identity* test $\{\mathscr{I}_\mathrm{A}\}$ such that

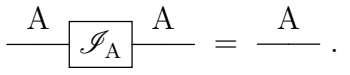


(3.2)



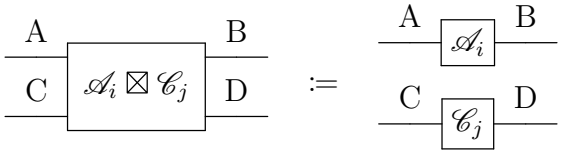
(3.3)

for any system B and C and any test $\{\mathscr{A}_i\}_{i\in\mathsf{X}}, \{\mathscr{B}_j\}_{j\in\mathsf{Y}}$ from A to B and C to A respectively. Diagrammatically, the identity test will be represented simply by a wire, according to the fact that they can be understood as doing nothing on the system at hand.

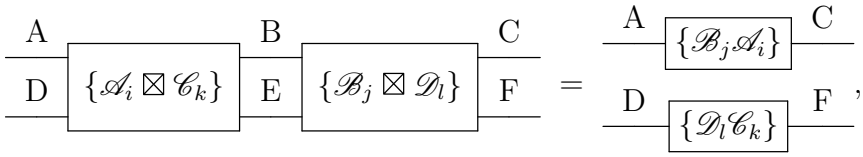


(3.4)

Another operation that can be performed on tests is *parallel composition*. Given two systems A and B we call AB the composite system of A and B. Then, if $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ and $\{\mathscr{C}_j\}_{j\in\mathsf{Y}}$ are two tests, we define their parallel composition as the test $\{\mathscr{A}_i \boxtimes \mathscr{C}_j\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$, with outcome space given by the cartesian product $\mathsf{X}\times\mathsf{Y}$. Diagrammatically, the new test is represented by vertical juxtaposition of the composing tests



The symbol $\boxtimes$ is used in order to avoid confusion with the symbol $\otimes$, which denotes the tensor product of vector spaces. As for sequential composition, also parallel composition is required to be an associative operation. Moreover, the parallel composition operation commutes with the sequential one, namely, for all tests $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$, $\{\mathscr{B}_j\}_{j\in\mathsf{Y}}$, $\{\mathscr{C}_k\}_{k\in\mathsf{Z}}$, $\{\mathscr{D}_l\}_{l\in\mathsf{L}}$, the following equation holds

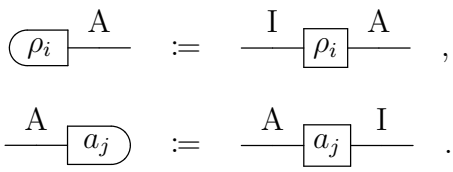


(3.5)

or, in formula

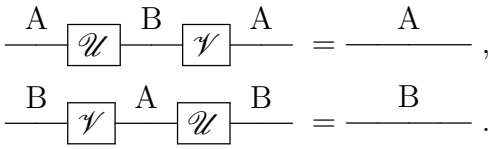
$$(\mathscr{B}_j \boxtimes \mathscr{D}_l)(\mathscr{A}_i \boxtimes \mathscr{C}_k) = (\mathscr{B}_j\mathscr{A}_i) \boxtimes (\mathscr{D}_l\mathscr{C}_k). \tag{3.6}$$

There is a special kind of system, the *trivial system* I, satisfying AI = IA = A for every system A. Tests with I as input system and A as the output one are called *preparation tests* of A, while tests with input system A and I as output are named *observation tests* of A. The events of a preparation test $\{\rho_i\}_{i\in\mathsf{X}}$ and of an observation test $\{a_j\}_{j\in\mathsf{Y}}$ are represented through the following diagrams



In the following we will always use Greek letters to denote preparation tests and Latin letters for the observation tests. Preparation and observation events will also be denoted by using round brackets, respectively $|\rho)_{\mathrm{A}}$ and $(a|_{\mathrm{A}}$, and we will not make explicit the system whenever it is clear from the context.

We say that $\mathscr{U} \in \mathsf{Event}(\mathrm{A} \to \mathrm{B})$ is *reversible* if there exists $\mathscr{V} \in \mathsf{Event}(\mathrm{B} \to \mathrm{A})$ such that $\mathscr{V}\mathscr{U} = \mathscr{I}_{\mathrm{A}}$ and $\mathscr{U}\mathscr{V} = \mathscr{I}_{\mathrm{B}}$, namely

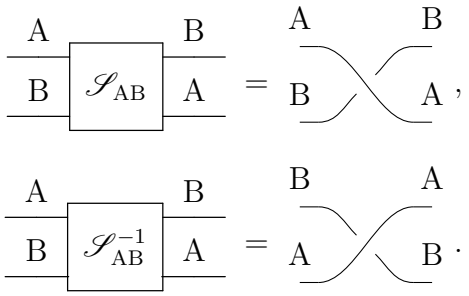


(3.7)

(3.8)

Accordingly, two systems A and B are called *operationally equivalent* if there exists a reversible transformation $\mathscr{U} \in \mathsf{Event}(\mathrm{A} \to \mathrm{B})$. Finally, we say that a transformation $\mathscr{V} \in \mathsf{Event}(\mathrm{A} \to \mathrm{B})$ is *left-reversible* if there exists $\mathscr{V}^{-1} \in \mathsf{Event}(\mathrm{B} \to \mathrm{A})$ such that $\mathscr{V}^{-1}\mathscr{V} = \mathscr{I}_{\mathrm{A}}$ (only the equation (3.7)). We also notice that both reversible and left-reversible operations must be necessarily deterministic.

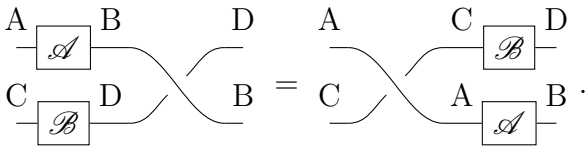
For any pair of systems, we want their agents to be allowed to exchange their systems. This requirement is captured by the notion of *braiding*, that is a family of reversible transformations $\mathscr{S}_{\mathrm{AB}}$, defined for any pair of systems of the theory, and denoted as follows



(3.9)

(3.10)

Generally, $\mathscr{S}_{\mathrm{AB}}$ and $\mathscr{S}_{\mathrm{AB}}^{-1}$ are different transformations. When they are equal, e.g. in the case of QT, where $\mathscr{S}_{\mathrm{AB}}$ is represented by the swap operator[2] , the theory is called symmetric. These transformations must obey a sliding property, which asserts that two agents can equivalently perform their transformations and then exchange their output systems, or first exchange their input systems and then perform the transformations. In diagrams



(3.11)

Finally, a *circuit* is a diagram representing an arbitrary test that is obtained by sequential and parallel composition of other tests. We say that a circuit is *closed* when the input and output systems are both the trivial one, namely, when it starts with a preparation test and it ends with an observation test.

Now that we have outlined all the essential elements, we are in the position to give the definition of operational theory.

Definition 3.1.1. *An* Operational Theory *is defined by a collection of systems, that is closed under composition, and a collection of tests (including all the family of tests $\mathscr{S}_{\mathrm{AB}}$ for any* A *and* B*) closed under a pair of associative operations of sequential and parallel composition.*

3.1.2 The probabilistic structure

Any closed circuit clearly represents a complete experimental apparatus. Since a physical theory is required to predict the probabilites with which outcomes occurr, we impose that closed circuits correspond to joint probaiblity distributions on the outcomes, conditioned by the tests of which the circuits are made.

Definition 3.1.2. *An* Operational Probabilistic Theory *is an operational theory where any closed circuit or, equivalently, any test from the trivial system to*

[2]More precisely, in QT $\mathscr{S}_{\mathrm{AB}}$ is the quantum channel whose Kraus operator is given by the unitary operator $S : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{A}}$ such that $S|\psi\rangle \otimes |\phi\rangle = |\phi\rangle \otimes |\psi\rangle$ for all $\psi \in \mathcal{H}_{\mathrm{A}}$ and $\phi \in \mathcal{H}_{\mathrm{B}}$.

istelf, is given by a joint probability distribution conditioned by the tests building the circuit.

The simplest example of a closed circuit is given by a preparation test $\{\rho_i\}_{i\in\mathsf{X}}$ sequentially followed by an observation test $\{a_j\}_{j\in\mathsf{Y}}$:

$$\{p(i,j|\{\rho_i\},\{a_j\})\} \coloneqq \left\{ \rho_i \xrightarrow{\mathrm{A}} a_j \right\}, \tag{3.12}$$

with $\sum_{i,j} p(i,j) = 1$. Thus, one has a joint probability distribution, which is conditioned by the chosen tests $\{\rho_i\}_{i\in\mathsf{X}}$ and $\{a_j\}_{j\in\mathsf{Y}}$. From now on we will simply omit this dependence whenever the tests are clear from the context. The probability associated with the closed circuit where a preparation ρ_i is followed by an observation a_j will also be denoted by a pairing, $p(i,j) = (a_j|\rho_i)$.

Moreover, compound tests from the trivial system to istelf are independent, namely, the joint probability distribution is simply given by the product of the probability distributions of the composing tests. This means that, if $\{\rho_i\}i \in \mathsf{X}$ and $\{\sigma_k\}_{k\in\mathsf{Z}}$ are two preparation tests, $\{a_j\}_{j\in\mathsf{Y}}$ and $\{b_l\}_{l\in\mathsf{L}}$ are observation tests, then

$$\begin{array}{l} \rho_i \xrightarrow{\mathrm{A}} a_j \\ \sigma_i \xrightarrow{\mathrm{A}} b_l \end{array} = p(i,j)q(k,l). \tag{3.13}$$

Given any system A of an OPT, one can define an equivalence relation on the set of preparation events by declaring that

$$\rho \sim \sigma \iff (a|\rho) = (a|\sigma), \qquad \forall a \in \mathsf{Event}(\mathrm{A} \to \mathrm{I}). \tag{3.14}$$

The set of equivalence classes with respect to this relation is called the set of *states* of system A, and it is denoted by $\mathsf{St}(\mathrm{A})$. We then set

$$\mathsf{St}(\mathrm{A}) := \mathsf{Event}(\mathrm{I} \to \mathrm{A})/\sim \ . \tag{3.15}$$

Similarly, one can define the set of *effects* as the set of equivalence classes of the observation events such that $(a|\rho) = (b|\rho)$ for any preparation event ρ, so that we can analogously define

$$\mathsf{Eff}(\mathrm{A}) := \mathsf{Event}(\mathrm{A} \to \mathrm{I})/\sim \ . \tag{3.16}$$

By definition, it is clear that the set of states is separating[3] for that of effects, and viceversa. The sets of deterministic states and effects will be denoted by $\mathsf{St}_1(\mathrm{A})$ and $\mathsf{Eff}_1(\mathrm{A})$ respectively.

Consider an arbitrary state of the theory, say ρ, and an observation test $\{a_j\}_{j\in\mathsf{Y}}$. Upon marginalisation over the observation test, one can define the *preparation probability conditioned by the test* $\{a_j\}_{j\in\mathsf{Y}}$ as

$$p(\rho|\{a_j\}) \coloneqq \sum_j (a_j|\rho). \tag{3.17}$$

[3]Generally, let S be a set, S^* be the set of real functionals on S and $\mathsf{G} \subseteq \mathsf{S}^*$. We say that G is *separating* for S if for any $s_1, s_2 \in \mathsf{S}$ there exists $g \in \mathsf{G}$ such that $g(s_1) \neq g(s_2)$.

Generally, one has that $p(\rho|\{a_j\}) \leq 1$, unless the preparation test $\{\rho_i\}_{i\in\mathsf{X}}$ is the singleton, i.e. the state is deterministic. Moreover, it is worth noticing that it can also depend on the observation test we are marginalising over.

Given the probabilistic structure, states can be seen as functionals on the set of effects and viceversa, and then one can consider linear combinations of them, thus defining two linear spaces, $\mathsf{St}(\mathrm{A})_{\mathbb{R}}$ and $\mathsf{Eff}(\mathrm{A})_{\mathbb{R}}$, which are dual to each other assuming that they are finite-dimensional

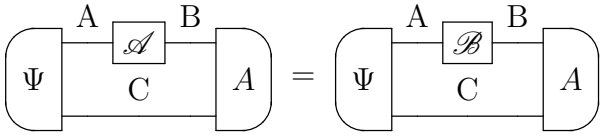
$$\mathsf{St}(\mathrm{A})_{\mathbb{R}} = [\mathsf{Eff}(\mathrm{A})_{\mathbb{R}}]^*, \qquad \mathsf{Eff}(\mathrm{A})_{\mathbb{R}} = [\mathsf{St}(\mathrm{A})_{\mathbb{R}}]^*. \tag{3.18}$$

The *size* D_{A} of a given system A is simply the dimension of the linear space $\mathsf{St}_{\mathbb{R}}(\mathrm{A})$, that is equal to that of $\mathsf{Eff}_{\mathbb{R}}(\mathrm{A})$ in the finite-dimensional case.

$$D_{\mathrm{A}} := \dim[\mathsf{St}_{\mathbb{R}}(\mathrm{A})]. \tag{3.19}$$

A transformation event from system A to system B induces a linear map from $\mathsf{St}_{\mathbb{R}}(\mathrm{AC})$ to $\mathsf{St}_{\mathbb{R}}(\mathrm{BC})$ for any ancillary system C. Also the set of transformation events can be endowed with an equivalence relation.

Definition 3.1.3. *Given two events $\mathscr{A}$ and $\mathscr{B}$, we say that they are* operationally equivalent*, and write $\mathscr{A} \sim \mathscr{B}$, if the following identity holds*



(3.20)

for any $\Psi \in \mathsf{St}(\mathrm{AC})$, $A \in \mathsf{Eff}(\mathrm{BC})$ and any ancillary system C.

In other words, two transformation events are operationally equivalent if they induce the same linear map for any ancillary system C, and it is clear that this is an equivalence relation on the set $\mathsf{Event}(\mathrm{A} \to \mathrm{B})$ for any pair of systems A and B. We then denote the set of all the equivalence classes of events with $\mathsf{Transf}(\mathrm{A} \to \mathrm{B})$, whose elements are simply called *transformations*. Namely, we formally set

$$\mathsf{Transf}(\mathrm{A} \to \mathrm{B}) := \mathsf{Event}(\mathrm{A} \to \mathrm{B})/\sim \tag{3.21}$$

Notice that this definition boils down to the case of states and effects when we set $\mathrm{A} = \mathrm{I}$ and $\mathrm{B} = \mathrm{I}$ respectvely, i.e. $\mathsf{St}(\mathrm{B}) = \mathsf{Transf}(\mathrm{I} \to \mathrm{B})$ and $\mathsf{Eff}(\mathrm{A}) = \mathsf{Transf}(\mathrm{A} \to \mathrm{I})$. The set of deterministic transformations will be denoted by $\mathsf{Transf}_1(\mathrm{A} \to \mathrm{B})$. Given a transformation $\mathscr{A} \in \mathsf{Transf}(\mathrm{AC} \to \mathrm{BC})$ is said to be *local from* A *to* B if it is of the form $\mathscr{A} = \mathscr{T} \boxtimes \mathscr{I}_{\mathrm{C}}$, for some $\mathscr{T} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$. Similarly, a transformation $\mathscr{A} \in \mathsf{Transf}(\mathrm{CA} \to \mathrm{CB})$ is said to be *local from* A *to* B if it is of the form $\mathscr{A} = \mathscr{I}_{\mathrm{C}} \boxtimes \mathscr{T}$, for some $\mathscr{T} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$.

The linear space $\mathsf{St}_{\mathbb{R}}(\mathrm{A})$ can be endowed with a metric structure by means of the following norm, which has an operational meaning related to optimal discrimination schemes [43] (see also remark 2.2.2).

Definition 3.1.4 (Operational norm)**.** *The norm of an element $\rho \in \mathsf{St}(\mathrm{A})_{\mathbb{R}}$ is defined as*

$$\|\rho\|_{\mathrm{op}} \coloneqq \sup_{\{a_0,a_1\}\subseteq \mathsf{Eff}(\mathrm{A})} (a_0 - a_1|\rho),$$

where $\{a_0, a_1\}$ is any binary observation test.

For the proof of the fact that this is, indeed, a norm see also [57]. This norm satisfies a monotonicity property, as stated in the following lemma.

Lemma 3.1.1 (Monotonicity of the operational norm)**.** *For any $\delta \in \mathsf{St}_{\mathbb{R}}(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A} \to \mathrm{B})$ the following inequality holds*

$$\|\mathscr{C}\delta\|_{\mathrm{op}} \leq \|\delta\|_{\mathrm{op}}, \tag{3.22}$$

with the equality holding if $\mathscr{C}$ is reversible.

Now we set up some terminology and we introduce pure and mixed states, as well as the definition of state dilation.

Definition 3.1.5 (Refinement and convex refinement of an event)**.** *Let $\mathscr{C} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$.*

1. *A* refinement *of $\mathscr{C}$ is given by a collection of events $\{\mathscr{B}_j\}_{j\in\mathsf{Y}} \subseteq \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ such that there exists a test $\{\mathscr{B}_i\}_{i\in\mathsf{X}}$ with $\mathsf{Y} \subseteq \mathsf{X}$ and $\mathscr{C} = \sum_{j\in\mathsf{Y}} \mathscr{B}_j$. We denote by $\mathsf{Ref}(\mathscr{C})$ the set of all the refinements of $\mathscr{C}$. We say that a refinement $\{\mathscr{D}_i\}_{i\in\mathsf{Y}}$ is trivial if $\mathscr{D}_i = \lambda_i\mathscr{C}$, $\lambda_i \in [0, 1]$ for every $i \in \mathsf{Y}$. Conversely, $\mathscr{C}$ is called the coarse-graining of the events $\{\mathscr{D}_i\}_{i\in\mathsf{Y}}$.*

2. *A* convex refinement *or* decomposition *of $\mathscr{C}$ is given by a collection of pairs $\{(p_j, \mathscr{B}_j)\}_{j\in\mathsf{Y}} \subseteq \mathbb{R} \times \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ where $\{p_j\}_{j\in\mathsf{Y}}$ is a probability distribution and $\mathscr{B}_j$ are events, such that there exist tests $\{\mathscr{B}_i^{(j)}\}_{i\in\mathsf{X}}$, with $\mathscr{B}_{i_0}^{(j)} = \mathscr{B}_j$ for all $j \in \mathsf{Y}$, for which the collection $\{p_j\mathscr{B}_i^{(j)}\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$ is a legitimate conditional test, and $\mathscr{C} = \sum_{j\in\mathsf{Y}} p_j\mathscr{B}_j$. We say that a convex refinement is trivial if $\mathscr{B}_j = \mathscr{C}$ for any $j \in \mathsf{Y}$.*

Definition 3.1.6. *Given two events $\mathscr{C}, \mathscr{D} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ we say that $\mathscr{D}$ refines $\mathscr{C}$, and write $\mathscr{D} \prec \mathscr{C}$, if there exists a refinement $\{\mathscr{D}_i\}_{i\in\mathsf{X}}$ of $\mathscr{C}$ such that $\mathscr{D} \in \{\mathscr{D}_i\}_{i\in\mathsf{X}}$.*

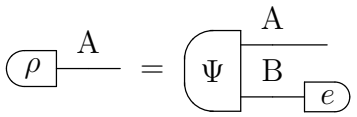
Definition 3.1.7 (Atomic, refinable, and extremal events)**.** *An event $\mathscr{C}$ is called atomic if it admits only trivial refinements. An event is refinable if it is not atomic. Finally, an event is extremal if it admits only trivial convex refinements.*

The notion of convex refinement and extremal events can be used in order to define the notions of *pure* and *mixed* states.

Definition 3.1.8 (Pure and mixed states)**.** $\rho \in \mathsf{St}(\mathrm{A})$ *is called pure if it is extremal and deterministic, and we will denote by* $\mathsf{PurSt}(\mathrm{A})$ *the set of all the pure states of system* A*.* $\rho \in \mathsf{St}(\mathrm{A})$ *is said to be mixed if it is neither atomic nor pure, and if* $\mathsf{Ref}(\rho) = \mathsf{St}(\mathrm{A})$*, then it is said to be* completely mixed*.*

In the literature, the terms "atomic" and "pure" are often used interchangeably. In quantum theory, the atomic states are all and only those of the form $\lambda \left|\phi\right\rangle \left\langle\phi\right|$, with $\lambda \in [0, 1]$, therefore, they are all proportional to a pure state. However, as pointed out in [44], these two notions must be kept distinct, since there exist state spaces with pure states that are not necessarily atomic. The correspondence is meaningful in those theories in which every state is proportional to a deterministic one. Moreover, in the state spaces exhibited in [44] there can be extremal states that are not deterministic, so that they are examples of states spaces which show that the set of extremal states can generally be bigger that the set of the pure ones.

Definition 3.1.9 (State dilation)**.** *Let* $\rho \in \mathsf{St}(\mathrm{A})$ *and* $\Psi \in \mathsf{St}(\mathrm{AB})$*. We say that* Ψ *is a dilation of* ρ *if there exists a deterministic effect* $e \in \mathsf{Eff}(\mathrm{B})$ *such that*



We denote by D_ρ *the set of all dilations of the state* ρ*. If* Ψ *is also pure, then we say that it is a purification of* ρ *and* B *is called the purifying system. Finally, we denote by* P_ρ *the set of all the purifications of* ρ*.*

Trivially one has that $P_\rho \subseteq D_\rho$. Moreover, if $\Omega \in D_\rho$, then one has $D_\Omega \subseteq D_\rho$, as it can be easily verified.

The notion of entangled state is easily generalised to OPTs by substituting the tensor product with the operation of parallel composition.

Definition 3.1.10 (Separable and entangled states)**.** *Consider a bipartite system* AB *and let* $\Sigma \in \mathsf{St}(\mathrm{AB})$*. We say that* Σ *is a* separable *state if there exist* $\{\rho_i\}_{i\in\mathsf{X}} \subseteq \mathsf{St}(\mathrm{A})$ *and* $\{\sigma_i\}_{i\in\mathsf{X}} \subseteq \mathsf{St}(\mathrm{B})$ *such that*

$$|\Sigma)_{\mathrm{AB}} = \sum_{i\in\mathsf{X}} |\rho_i)_{\mathrm{A}} \boxtimes |\sigma_i)_{\mathrm{AB}}.$$

By negation, entangled *states are those that are non-separable. States of the form* $|\rho)_{\mathrm{A}} \boxtimes |\sigma)_{\mathrm{B}}$ *are called* factorised*.*

3.1.3 Different degrees of locality: n-local discriminability

As we deduce from the definition of operationally equivalent transformation events, definition 3.1.3, in order to distinguish two transformations belonging to distinct equivalence classes, it is necessary to test them on *all* possible states and effects with *any* conceivable ancilla. A way of reconciling this holistic trait

of the OPT framework with a reductionist scientific approach is to say that a theory is, in some sense, local [58]. Here we consider the following definition [59, 60].

Definition 3.1.11 (n-local discriminability)**.** *Let $n \leq m$. The effects obtained as a conic combination of the parallel composition of effects $a_1, \ldots, a_n$ where a_j is k_j-partite with $k_j \leq n$ for all values of j, are separating for m-partite states.*

Clearly, a theory that is n-locally discriminable, it is also n'-locally discriminable for any $n' \geq n$. So that it makes sense to define a *strictly* n-locally discriminable theory as one that is n'-locally discriminable only for $n' \geq n$.

Actually, the case with $n = 1$, simply known as *local discriminability* [43, 17, 57], was already considered in the literature on GPTs and OPTs, and it is a property of QT. This feature is equivalent to another property, named local tomograhy, or tomographic locality, corresponding to the possibility of state reconstruction via local effects [61, 47, 45, 62]), and it also has remarkable consequences that are collected in the following proposition

Proposition 3.1.1. *In any OPT the following facts are equivalent:*

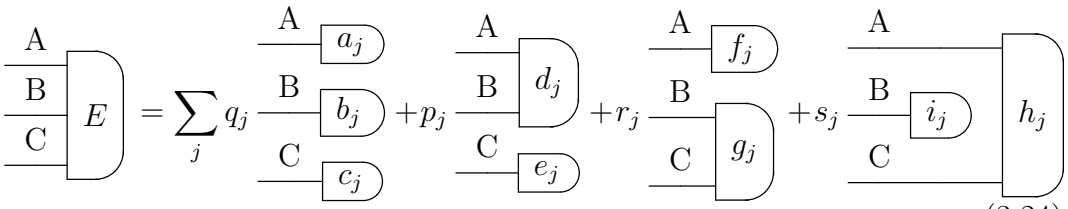
1. *local discriminability holds;*
2. *For any pair of systems* A *and* B*, given $\mathscr{A}$, $\mathscr{B} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$, if $\mathscr{A}|\rho)_{\mathrm{A}} = \mathscr{B}|\rho)_{\mathrm{A}}$ for all $\rho \in \mathsf{St}(\mathrm{A})$ then $\mathscr{A} = \mathscr{B}$;*
3. *$D_{\mathrm{AB}} = D_{\mathrm{A}} D_{\mathrm{B}}$.*

Notice that the tensor product rule with which systems are composed in QT trivially implies $D_{\mathrm{AB}} = D_{\mathrm{A}} D_{\mathrm{B}}$. In general, the linear space spanned by the set of states of a bipartite system can be decomposed in a local and a non-local part as $\mathsf{St}_{\mathbb{R}}(\mathrm{AB}) := [\mathsf{St}(\mathrm{A}) \boxtimes \mathsf{St}(\mathrm{B})] \oplus \mathsf{St}^{\mathrm{NL}}_{\mathbb{R}}(\mathrm{AB})$, where $\mathsf{St}^{\mathrm{NL}}_{\mathbb{R}}(\mathrm{AB})$ defines the subspace containing the non-local contributions to states. But in the presence of local discriminability, one simply has $\mathsf{St}_{\mathbb{R}}(\mathrm{AB}) := \mathsf{St}(\mathrm{A}) \boxtimes \mathsf{St}(\mathrm{B})$, and the operation of parallel composition $\boxtimes$ can actually be understood as a tensor product $\otimes$. The second condition, also called *local process tomography*, is quite remarkable. Indeed, as we have already said, from the definition of operationally equivalent events, in order to actually distinguish two transformations it might be necessary to consider all the possible ways in which our local system is correlated with any ancilla. However, the condition of item 2, also known as *local process tomography*, simplifies the task of discriminating $\mathscr{A}$ and $\mathscr{B}$ to evaluating their action on states of the local system A only. In mathematical terms, given $\mathscr{A}, \mathscr{B} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$,

$$\mathscr{A} = \mathscr{B} \iff \mathscr{A}|\rho)_{\mathrm{A}} = \mathscr{B}|\rho)_{\mathrm{A}} \qquad \forall \rho \in \mathsf{St}(\mathrm{A}), \tag{3.23}$$

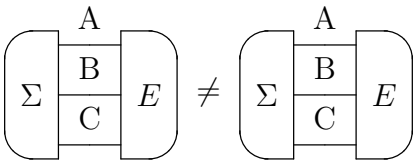
whenever the theory satisfies local discriminability.

Another case that has been extensively studied in the literature is that of strictly bilocal theories, corresponding to $n = 2$. In such a case there exist two distinct states $\Psi, \Phi \in \mathsf{St}(\mathrm{AB})$ that: i) are equal on all the local effects; ii) any m-partite state, with $m \geq 3$, can be distinguished from any other by an effect which is a conic combination consisting of the parallel composition of, at most, 2-partite effects. Namely, whenever $\Gamma, \Sigma \in \mathsf{St}(\mathrm{ABC})$ are such that $\Sigma \neq \Gamma$ there exists an effect $E \in \mathsf{Eff}(\mathrm{ABC})$ of the following form



(3.24)

with all $q_j, p_j, r_j, s_j \geq 0$ and such that



(3.25)

Given the conic decomposition of a 2-local effect as in (3.24), one has the following relation for system sizes [59, 60].

$$D_{\mathrm{AB}} \geq D_{\mathrm{A}} D_{\mathrm{B}}, \tag{3.26}$$

$$D_{\mathrm{ABC}} = D_{\mathrm{A}} D_{\mathrm{B}} D_{\mathrm{C}} + D_{\mathrm{A}} \Delta_{\mathrm{BC}} + D_{\mathrm{B}} \Delta_{\mathrm{AC}} + D_{\mathrm{C}} \Delta_{\mathrm{BC}}, \tag{3.27}$$

Where $\Delta_{\mathrm{AB}} := D_{\mathrm{AB}} - D_{\mathrm{A}} D_{\mathrm{B}}$. In a *strictly bilocal theory*, inequality in (3.26) must hold without equality for at least a pair of systems. Real quantum theory, fermionic theory, that will be reviewed in later in this chapter, and bilocal classical theory, that wil be discussed in subsection 3.2.3, are examples of strictly bilocal theories.

3.2 Special instances of the framework: Quantum Theory, Fermionic Theory and classical theories

Up to now we have presented a summary of the structure of the framework, in particular we have declaired what are the primitive notions of an OPT, along with the possible operations that we naturally admit thereon, and we have seen how probabilities define a linear structure on the sets of events. CT and QT and FT are two particular instances of this framework.

3.2.1 The axioms of Quantum Theory

In section 2.2 we revised the mathematical structure of QT, in its Hilbert space formulation. Such a description can also be derived in the OPT framework described in the previous section, in an axiomatic fashion inspired by the insights emerging from the experience of quantum information theory. In [17] the Hilbert space formalism was derived from the following set of axioms.

Axiom 1 (Weak causality)**.** *Let* $\rho \in \mathsf{St}(\mathrm{A})$ *and* $\{a_i\}_{i\in\mathsf{X}}$, $\{b_j\}_{j\in\mathsf{Y}}$ *be two observation tests. Then an OPT is said to be* weakly causal *if the following holds*

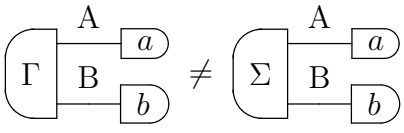
$$\sum_{i\in\mathsf{X}} p(\rho|\{a_i\}) = \sum_{j\in\mathsf{Y}} p(\rho|\{b_j\}). \tag{3.28}$$

This is a common notion of causality considered in the literature and it expresses a *no-signalling from the future* principle, namely it asserts that preparation probabilities are actually independent of the chosen observation test; it is also equivalent to state the uniqueness of the deterministic effect. It is clear that this is a property of QT, given that any observation test, represented by a POVM $\{E_i\}$, sums to the identity operator I. A remarkable fact is that, in a bipartite scenario, this also implies a no-signalling in space, unless the two parties decide to exchange their systems [43]. A simple example of a non-causal OPT can be obtained as the dual theory of a causal one (such as CT), namely that obtained by exchanging the role of states and effects. Indeed, in this case there can be more than one deterministic effect, therefore weak causality is violated. Other more elaborated examples can be found in [57, 53]

Axiom 2 (Perfect discriminability)**.** *Any state that is not completely mixed can be perfectly distinguished from some other state.*

Axiom 3 (Ideal compression)**.** *For any* A *and* $\rho \in \mathsf{St}_1(\mathrm{A})$ *there exist a system* C *of size* $D_{\mathrm{C}} \leq D_{\mathrm{A}}$ *and a pair of maps* $\mathscr{E} \in \mathsf{Transf}(\mathrm{A}\to\mathrm{C})$, $\mathscr{D} \in \mathsf{Transf}(\mathrm{C}\to\mathrm{A})$ *such that* $\mathscr{D}\mathscr{E}(\sigma_i) = \sigma_i$ *for any* $\{\sigma_i\} \subseteq \mathsf{Ref}(\rho)$.

Axiom 4 (Local discriminability)**.** *For any pair of systems* A *and* B*, two states* Σ *and* Γ *in* $\mathsf{St}(\mathrm{AB})$ *are different iff there exists a pair of effects* $a \in \mathsf{Eff}(\mathrm{A})$ *and* $b \in \mathsf{Eff}(\mathrm{B})$ *such that*



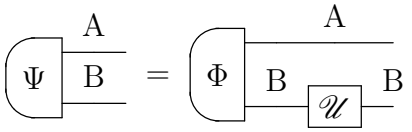
(3.29)

We have already discussed this feature at the end of the previous section, together with some of its consequences. Here we have made it explicit in mathematical terms.

Axiom 5 (atomicity of sequential composition)**.** *If* $\mathscr{A} \in \mathsf{Transf}(\mathrm{A}\to\mathrm{B})$ *and* $\mathscr{B} \in \mathsf{Transf}(\mathrm{B}\to\mathrm{C})$ *are atomic transformations, then so is* $\mathscr{C} := \mathscr{B}\mathscr{A} \in \mathsf{Transf}(\mathrm{A}\to\mathrm{C})$.

Axiom 6 (Existence of purification)**.** *We say that an OPT satisfies purification if for any* $\rho \in \mathsf{St}(\mathrm{A})$ *the set of purifications of* ρ *in not empty, i.e.* $P_\rho \neq \emptyset$.

Axiom 7 (Essential uniqueness of purification)**.** *We say that an OPT satisfies essential uniqueness of purification if, for any* $\rho \in \mathsf{St}(\mathrm{A})$ *such that* $P_\rho \neq \emptyset$, $\forall \Phi, \Psi \in P_\rho$ *with* $\Psi, \Phi \in \mathsf{St}(\mathrm{AB})$*, there exists a reversible transformation* $\mathscr{U}$ *such that*



(3.30)

Axioms 6 and 7 were formerly introduced in [43] and actually they were stated as a unique one; together, they constitute the formalisation of a characteristic trait of quantum theory, recalled in theorem 2.2.1. Axioms 6 states that any state can be simulated as the marginal of a pure state that lies on a larger system. Axiom 7 tells that the purifying state Φ is unique up to unitary quantum operations on the ancillary system E. CT satisfies the first five axioms listed above, except for axiom 6, that singles out the quantumness in the axiomatisation of [43].

Remark 3.2.1. A separate statement for these two properties is mandatory, since, as pointed out in [44], they are actually independent features, in the sense that there exist theories satisfying axiom 6 but not 7 and viceversa; in poor words, none of them implies the other.

In QT, not only sequential composition is atomicity-preserving, but also the parallel composition operation. In particular, composition of pure states trivially yields another (factorised) pure state. The latter property plays an important role in our treatment, and it is worthy of a statement in its own right.

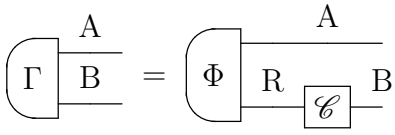
Property 1 (Atomicity and purity of parallel composition of states)**.** *We say that an OPT satisfies:*

1. *atomicity of parallel composition of states whenever, for any pair of systems* A *and* B*, given two atomic states* $\phi \in \mathsf{St}(\mathrm{A})$ *and* $\psi \in \mathsf{St}(\mathrm{B})$*, their parallel composition* $\phi \boxtimes \psi$ *is atomic. Accordingly, we say that the parallel composition is* atomicity-preserving*;*

2. *purity of parallel composition of states whenever, for any pair of systems* A *and* B*,* $\phi \in \mathsf{PurSt}(\mathrm{A})$ *and* $\psi \in \mathsf{PurSt}(\mathrm{B})$ *implies* $\phi \boxtimes \psi \in \mathsf{PurSt}(\mathrm{AB})$*. Accordingly, we say that the parallel composition is* purity-preserving*.*

When Axiom 7 is used along with item 2 of property 1 one can prove [43] that any two purifications of the same state are always connected by a suitable channel. When also 6 holds, then any dilation can be obtained by a fixed purification, as stated in the following propositions

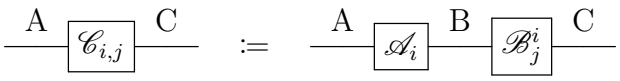
Proposition 3.2.1. *Let $\rho \in \mathsf{St}(\mathrm{A})$ be a state of an OPT satisfying existence and uniqueness of purification (axioms 6 and 7) and item 2 of property 1. Then, given a purification $\Phi \in \mathsf{PurSt}(\mathrm{AR})$, for any dilation $\Gamma \in \mathsf{St}(\mathrm{AB})$ there exists a channel $\mathscr{C} \in \mathsf{Transf}(\mathrm{R} \to \mathrm{B})$ such that*



(3.31)

The notion of causality expressed in axiom 1 is related to a stronger property satisfied by quantum theory, the possibility of *conditioning*, that is defined as the possibility of choosing a test in a given collection conditioned on the outcome of another test.

Property 2 (Conditioning)**.** *An OPT satisfies strong causality if for every test $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ and every collection of tests $\{\mathscr{B}^i_j\}_{j\in\mathsf{Y}}$ labelled by $j \in \mathsf{Y}$, the collection of events $\{\mathscr{C}_{i,j}\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$ with*



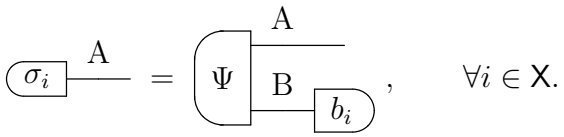
(3.32)

is a test of the theory.

This feature represents a stronger form of causality, since it implies uniqueness of the deterministic effect [57, 43], whereby it is also known as *strong causality*. Indeed, if by contradiction there exist more than one deterministic effect, with conditioning one would be able to construct a test that violates the rule of probability. In strongly causal theories, one can always assume that any state is proportional to a deterministic one without introducing any sort of inconsistency (see [63]).

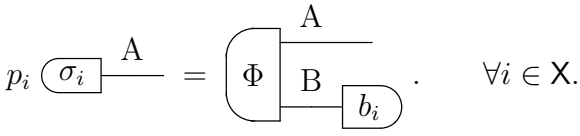
Another feature of QT is the so-called Steering property [64, 65], that can be generalised as follows for a generic OPT.

Property 3 (Steering)**.** *We say that an OPT satisfies steering if the following statement holds: Let $\rho \in \mathsf{St}(\mathrm{A})$ and $\{\sigma_i\}_{i\in\mathsf{X}} \subseteq \mathsf{St}(\mathrm{A})$ be a refinement of ρ. Then there exist a system B, a state $\Psi \in \mathsf{St}(\mathrm{AB})$, and an observation test $\{b_i\}_{i\in\mathsf{X}}$ such that*



$\forall i \in \mathsf{X}.$

Notice that the state Ψ in the statement must be a dilation of ρ, as one can easily verify upon summing over $i \in \mathsf{X}$. In QT, a stronger formulation of this property holds: given a state $\rho \in \mathsf{St}(\mathrm{A})$ and a purification $\Phi \in \mathsf{PurSt}(\mathrm{AB})$ of ρ, for any decomposition $\sum_{i\in\mathsf{X}} p_i\sigma_i$ of ρ there exists an observation test $\{b_i\}_{i\in\mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{B})$ such that



$\forall i \in \mathsf{X}.$

The stronger steering feature satisfied by quantum theory can actually be proven to hold in any OPT satisfying local discriminability, existence and uniqueness (up to reversible channels) of purification and perfect discriminability as axioms [57].

3.2.2 Review of Fermionic Theory

We now briefly review fermionic information theory [60, 66], which will be considered in chapter 5. The systems of the theory are made by local fermionic modes (LFMs). A LFM is the counterpart of the qubit in quantum theory, and can be thought of as a cell that can be either empty or occupied by a fermionic excitation. An L-LFMs system, denoted $\mathrm{L_F}$, is described by L fermionic fields φ_i, satisfying the canonical anticommutation rule (CAR)

$$\{\varphi_i, \varphi_j^\dagger\} = \delta_{ij} I, \quad \{\varphi_i, \varphi_j\} = 0 \quad \text{where } i, j = 1, \ldots, L.$$

With these fields one constructs the occupation number operators $\varphi_i^\dagger \varphi_i$, which can be easily proved to have only eigenvalues 0 and 1. The common eigenvector $|\Omega\rangle$ of the operators $\varphi_i^\dagger \varphi_i$, $i = 1, \ldots, L$ with eigenvalue 0 defines the vacuum state $|\Omega\rangle\langle\Omega|$ of $\mathrm{L_F}$, representing the state in which all the modes are not excited. The fermionic vacuum state in terms of the field operators is given by

$$|\Omega\rangle\langle\Omega| = \prod_{i=1}^{L} \varphi_i \varphi_i^\dagger. \tag{3.33}$$

By applying the operators $\varphi_i^\dagger$ to $|\Omega\rangle$ the corresponding i-th mode is excited and, by raising $|\Omega\rangle$ in all possible ways, we get the 2^L orthonormal vectors forming the Fock basis in the occupation number representation: a generic element of this basis is

$$|n_1, \ldots, n_L\rangle := (\varphi_1^\dagger)^{n_1} \ldots (\varphi_L^\dagger)^{n_L} |\Omega\rangle, \tag{3.34}$$

with $n_i = \{0, 1\}$ corresponding to the occupation number at the i-th site. The total occupation number of a basis vector is defined as $\sum_i n_i$. The linear span of these vectors corresponds to the antisymmetric Fock space $\mathcal{F}_L$ of dimension $d_{\mathcal{F}_L} = 2^L$. Notice that the Fock space $\mathcal{F}_L$ is isomorphic to the Hilbert space $\mathcal{H}_L$ of L qubits, by the trivial identification of the occupation number basis with the qubit computational basis. This correspondence lies at the basis of the Jordan-Wigner isomorphism [67, 68, 69] typically used in the literature to map fermionic systems to qubits systems and vice-versa. We recall here the

definition of the Jordan-Wigner map

$$\begin{aligned} &J_L(\varphi_i) = \left(\bigotimes_{l=1}^{i-1} \sigma_l^z\right) \otimes \sigma_i^- \otimes \left(\bigotimes_{k=i+1}^{L} I_k\right), \\ &J_L(\varphi_i^\dagger) = J_L(\varphi_i)^\dagger, \\ &J_L(XY) = J_L(X)J_L(Y), \\ &J_L(aX + bY) = aJ_L(X) + bJ_L(Y), \end{aligned} \tag{3.35}$$

with X, Y linear combinations of products of field operators on the L-LFMs, and where we used the standard notation for Pauli sigma operators. In the following we will drop the dependence on the number of LFMs in the Jordan-Wigner map, namely we will write $J(X)$ in place of $J_L(X)$, when it will be clear from the context. Notice that the Jordan-Wigner isomorphism is implicitly defined in Eq. (3.34), and, as such, it depends on the arbitrary ordering of the modes. All such representations are unitarily equivalent.

Differently from standard quantum systems, fermionic systems satisfy the parity superselection rule [70, 71, 72, 73, 60], namely, states commute with the parity operator.

$$P = \frac{1}{2}(I + \prod_{i=1}^{L}(\phi_i\phi_i^\dagger - \phi_i^\dagger\phi_i). \tag{3.36}$$

One can decompose the Fock space $\mathcal{F}_L$ of system $\mathrm{L_F}$ in the direct sum $\mathcal{F}_L = \mathcal{F}_L^e \oplus \mathcal{F}_L^o$, with $\mathcal{F}_L^e$ and $\mathcal{F}_L^o$ the spaces generated by vectors with even and odd total occupation number, respectively. The convex set of states $\mathsf{St}(\mathrm{L_F})$ corresponds, in the Jordan-Wigner representation, to the set of density matrices on $\mathcal{F}_L$ of the form $\rho = \rho_e + \rho_o$, with $\rho_e, \rho_o \geq 0$, $\mathrm{Tr}[\rho_o] + \mathrm{Tr}[\rho_e] \leq 1$ and with ρ_e and ρ_o having support on $\mathcal{F}_L^e$ and $\mathcal{F}_L^o$, respectively, and pure states are represented by rank one density operators. Any admissible fermionic state is represented by a density matrix given by a linear combination of products of an even number of field operators. Conversely, every linear combination of products of an even number of field operators that gives rise to a density matrix is an admissible state. Analogously, effects in the set $\mathsf{Eff}(\mathrm{L_F})$ are represented by positive operators on $\mathrm{L_F}$ of the form $a = a_e + a_o$, with a_e and a_o having support on $\mathcal{F}_L^e$ and $\mathcal{F}_L^o$, respectively. Notice that set of states and effects of system $\mathrm{L_F}$ have dimension

$$D_{\mathrm{L_F}} := \dim[\mathsf{St}_{\mathbb{R}}(\mathrm{L_F})] = d_{\mathcal{F}_L}^2/2 = 2^{2L-1}, \tag{3.37}$$

compared to the quantum case where the set of states and effects associated to the Hilbert space $\mathcal{H}_L$ of L qubits has dimension $d_{\mathcal{H}_L}^2 = 2^{2L}$.

Given two fermionic systems L_F and M_F, we introduce the composition of the two as the system made of $K \equiv L + M$ LFMs, denoted with the symbol $\mathrm{K_F} := \mathrm{L_F} \boxtimes \mathrm{M_F}$, or simply $\mathrm{K_F} := \mathrm{L_F M_F}$. We keep using the symbol $\boxtimes$ to distinguish the fermionic parallel composition rule from the quantum one, corresponding to the tensor product $\otimes$.

Given a state $\Psi \in \mathsf{St}(\mathrm{L_F M_F})$, one can discard the subsystem $\mathrm{M_F}$ and consider the marginal state, which we denote by $\sigma := \mathrm{Tr}^f_{\mathrm{M_F}}(\Psi)$. We use the symbol $\mathrm{Tr}^f_{\mathrm{M_F}}$ to denote the fermionic partial trace on the subsystem $\mathrm{M_F}$. This is computed by performing the following steps (see ref. [60] for further details): (i) drop all those terms in Ψ containing an odd number of field operators in any of the LFMs in $\mathrm{M_F}$; (ii) remove all the field operators corresponding to the LFMs in $\mathrm{M_F}$ from the remaining terms. The fermionic trace $\mathrm{Tr}^f(\rho)$ of a state $\rho \in \mathsf{St}(\mathrm{M_F})$ is then defined as a special case of the partial one, corresponding to the case in which $L = 0$.

Fermionic transformations in $\mathsf{Transf}(\mathrm{L_F} \to \mathrm{L_F})$ and effects are characterised according to the following proposition and its subseqeuent corollary.

Proposition 3.2.2 (Fermionic transformations)**.** *All the transformations in the set* $\mathsf{Transf}(\mathrm{L_F} \to \mathrm{L_F})$ *with Kraus operators being linear combinations of products of either an even number or an odd number of field operators are admissible fermionic transformations. Conversely, each admissible fermionic transformation in* $\mathsf{Transf}(\mathrm{L_F} \to \mathrm{L_F})$ *has Kraus operators being superpositions of products of either an even number or an odd number of field operators.*

Corollary 3.2.1 (Fermionic effects)**.** *Fermionic effects are positive operators bounded by the identity operator that are linear combinations of products of an even number of field operators. Conversely, every linear combination of products of an even number of field operators that is represented by a positive operator bounded by the identity is a fermionic effect.*

The corollary follows immediately from Proposition 3.2.2, since an effect A is obtained as a fermionic transformation $\mathscr{A}$ followed by the discard map, i.e. the trace. Thus

$$\mathrm{Tr}[\rho A] = \mathrm{Tr}[\mathscr{A}(\rho)] = \sum_i \mathrm{Tr}[K_i \rho K_i^\dagger] = \mathrm{Tr}[\rho \sum_i K_i^\dagger K_i],$$

namely $A = \sum_i K_i^\dagger K_i$. Having the polynomial K_i a definite parity (though not necessarily the same for every i), A is an even polynomial.

Remark 3.2.2. The description of FT given above, together with the parity superselection rule, is derived in [60] as a consequence of the following assumptions, concerning states and effects of fermionic systems and the local charactersation of fermionic transformations:

1. FT is causal;
2. the states of system $\mathrm{L_F}$ are represented by density matrices on the antisymmetric Fock space $\mathcal{F}_L$;
3. the transformations on $\mathrm{L_F}$ are represented by linear Hermitian preserving maps;

4. the map $\mathscr{A}_i$ with kraus operators $A_i := \varphi_i + \varphi_i^\dagger$ is an actual transformation of the theory;

5. let $I_L = \{1, \dots, L\}$ and $\mathrm{L_F}$ the system corresponding to L-LFMs. Transformations with Kraus operators in the algebra of field operators φ_i, $\varphi_i^\dagger$ with $i \in I \subset I_L$ are local on the subsystem of the LFMs associated with I;

6. local transformations on a system retain the same Kraus representation when other systems are added or discarded:

7. the pairing between states and effects is given by the Born rule $(a|\rho) := \mathrm{Tr}(a\rho)$;

8. on a single LFM the pairing with the deterministic effect is represented by $(e|\rho) := \mathrm{Tr}(\rho)$.

In particular, proposition 3.2.2 and corollary 3.2.1 follow by this set of assumptions, along with the fact that states are represented by linear combination of an even product of field operators, whence they commute with the parity operator P (equation (3.36)).

Finally, the set of transformations from $\mathrm{L_F}$ to $\mathrm{M_F}$ $\mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$, is given by completely positive maps from $\mathsf{St}(\mathrm{L_F})$ to $\mathsf{St}(\mathrm{M_F})$ in the Jordan-Wigner representation, and deterministic transformations, also called *channels*, from $\mathrm{L_F}$ to $\mathrm{M_F}$, correspond to trace-preserving completely positive maps. Like in quantum theory, any fermionic transformation $\mathscr{C} \in \mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$ can be expressed in Kraus form

$$\mathscr{C}(\rho) = \sum_i C_i \rho C_i^\dagger, \tag{3.38}$$

with deterministic transformations having Kraus operators $\{C_i\}$ such that $J(\sum_i C_i^\dagger C_i) = I_{\mathcal{H}_L}$, $I_{\mathcal{H}_L}$ denoting the identity operator on $\mathcal{H}_L$. For a map $\mathscr{C} \in \mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$ with Kraus operators $\{C_i\}$, we define its Jordan-Wigner representative $J(\mathscr{C})$ as the quantum map with Kraus operators $\{J(C_i)\}$. Now, given two transformations $\mathscr{C} \in \mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$ and $\mathscr{D} \in \mathsf{Transf}(\mathrm{K_F} \to \mathrm{N_F})$, we denote by $\mathscr{C} \boxtimes \mathscr{D} \in \mathsf{Transf}(\mathrm{L_F K_F} \to \mathrm{M_F N_F})$ the *parallel composition* of $\mathscr{C}$ and $\mathscr{D}$, with Kraus operators $\{C_i D_j\}$, where $\{C_i\}$ are Kraus operators for $\mathscr{C}$ and $\{D_j\}$ for $\mathscr{D}$. We observe that in the Jordan-Wigner representation one generally has

$$J_{L+K}(C_i D_j) \neq J_L(C_i) \otimes J_K(D_j), \tag{3.39}$$

$$J_{L+K}(\mathscr{C} \boxtimes \mathscr{D}) \neq J_L(\mathscr{C}) \otimes J_K(\mathscr{D}). \tag{3.40}$$

If $\mathscr{C}$ is a transformation in $\mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$, its extension to a composite system $\mathrm{L_F N_F}$, is given by $\mathscr{C} \boxtimes \mathscr{I}$, with $\mathscr{I}$ the identity map of system $\mathrm{N_F}$—whose Jordan-Wigner representative is the quantum identity map—and its

Kraus operators involve field operators on the $\mathrm{L_F}$ modes only. It is worth noting that, despite the Jordan-Wigner representative of this map is not necessarily of the form $J_L(\mathscr{C}) \otimes \mathscr{I}$, upon suitable choice of the ordering of the LFMs that defines the representation, one can always reduce to the case where, actually,

$$J_{L+N}(\mathscr{C} \boxtimes \mathscr{I}) = J_L(\mathscr{C}) \otimes \mathscr{I}. \tag{3.41}$$

As a special case of the above composition rule, one has $\rho \boxtimes \sigma = \rho\sigma \in \mathsf{St}(\mathrm{L_F M_F})$ for the parallel composition of states $\rho \in \mathsf{St}(\mathrm{L_F})$ and $\sigma \in \mathsf{St}(\mathrm{M_F})$, and similarly $a \boxtimes b = ab \in \mathsf{Eff}(\mathrm{L_F M_F})$ for the parallel composition of effects $a \in \mathsf{Eff}(\mathrm{L_F})$ and $b \in \mathsf{Eff}(\mathrm{M_F})$.

3.2.3 Classical Theories

We briefly described CT in the first chapter, where we recalled what are the mathematical representatives of states, effects and transformations. From such a description of CT we readily see that is satisfies the first five axioms 1-5 encompassed by QT. Indeed, weak causality follows by the uniqueness of the determinstic effect, and since the parallel composition is given by the tensor product, the size of the composite system is the product of the sizes of the composing systems, thus local discriminability holds (see proposition 3.1.1). As we have already observed, axiom 6, among those mentioned in subsection 3.2.1, is the sole that is not satisfied by CT, thus it can considered the genuinely quantum feature in such a formalisation.

Now, a common way of thinking about classicality, in the literature on GPTs and OPTs, is connected to the geometric structure of the set of states. Usually, a system is called "classical" when the pure states are jointly perfectly discriminable, according to the following definition.

Definition 3.2.1 (Joint perfect discriminability)**.** *Let* A *be a system of the theory. A set of states* $\{|\rho_i)_{\mathrm{A}}\}_{i=1}^n$ *is jointly perfectly discriminable if there exists an observation test* $\{(a_i|_{\mathrm{A}}\}$ *such that:*

$$(a_i|\rho_j)_{\mathrm{A}} \propto \delta_{ij}, \qquad \forall i, j \in \{1, \dots, n\}.$$

Remark 3.2.3. In [52], it has been proven that the notion of *full information without disturbance* (FIWD), a typical classical feature, indeed implies in full generality that all the systems of the theory must be classical. In [74], the notion of compatibility is studied and extended to strongly causal OPTs (i.e. satisfyting property 2), and another feature considered distinctively classical, *full compatibility*, is thoroughly scrutinised. In particolar, in addition to be equivalent to FIWD, it also implies that all the systems of the theory must be classical.

In CT, pure states, in addition to being jointly perfectly discriminable, they also feature another geometrical property: for any classical system, the pure states together with the null state are the vertices of a simplex. In this sense, CT is an example of a *simplicial theory*, according to the following definition.

Definition 3.2.2 (Simplicial theories)**.** *A* simplicial theory *is a finite dimensional OPT where the extremal states of every system* A *are the vertices of a* D_{A}*-simplex.*

In [59], simplicial OPTs have been studied thoroughly. It has been shown that any simplicial theory which is n-locally discriminable (definition 3.1.11) contains states that do not admit a purification, thus generalizing the well-known result that holds true for CT. Another consequence of the simpliciality condition expressed by this definition, is that the theory is necessarily weakly causal (axiom 1). Below we recall a lemma from [59] (precisely, lemma 1) on the characterisation of the parallel composition of states in simplicial theories satisfying n-local discriminability for some positive integer n, since it will be useful in chapter 6.

Lemma 3.2.1. *Consider a simplicial OPT staisfying n-local discriminability for some positive integer n. Then, for all systems* A, B *and non-null extremal states* $|k)_{\mathrm{AB}}$*, there exists a unique product of non-null extremal states* $|i_k)_{\mathrm{A}} \boxtimes |j_k)_{\mathrm{B}}$ *such that* $|k)_{\mathrm{AB}}$ *convexly refines* $|i_k)_{\mathrm{A}} \boxtimes |j_k)_{\mathrm{B}}$*.*

In CT, local discriminability and the uniqueness of the pure state decompositon entail that states on bipartite systems are all and only of the form (2.4), thus separable. On the one hand, the presence of entanglement is considered as one of the sharpest quantum feature, that definitely marks the departure of QT from CT. However, in [59] the authors studied simplicial theories that are n-local discriminable with $n \geq 2$, proving that, in such a case, simplicial OPTs admit the existence of bipartite entangled states according to definition 3.1.10. More precisely, one has the following theorem.

Theorem 3.2.1. *Consider a simplicial OPT. The following fact are equivalent:*

1. *the theory admits entangled states;*
2. *purity of parallel composition of states is violated;*
3. *the theory does not satisfy local discriminability.*

In particular, while in general it is always true that local discriminability implies that the compositional law is purity-preserving, the converse can generally be false, a trivial example being Fermionic Theory, which is bilocal.

As we have already said, simpliciality implies weak causality. For the present purposes, defining classical theories starting from simplicial ones is sufficient, therefore, we leave aside the problem of defining what a classical theory is in a non-causal setting, which is beyond our scope.

Definition 3.2.3 (Classical theories)**.** *A* weakly causal classical theory *is a simplicial theory where the pure states of every system are jointly perfectly discriminable.*

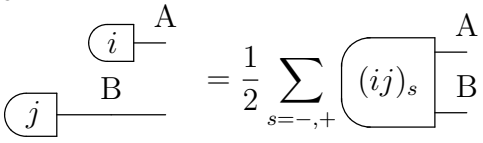
A classical theory called Bilocal Classical Theory, that violates purity of parallel composition of states and, according to theorem 3.2.1, admits of the presence of entangled states has been exhibited in [44]. BCT is a fully-fledged operational probabilistic theory that has been constructed with the aim of showing the independence of two fundamental notions: entanglement and complementarity. As its name suggests, BCT is a strictly bilocal theory, in the sense described in subsection 3.1.3, and as such it represents a good test to analyse the behaviour and the properties of the operational definition of information content that we will discuss in the next chapter. Here we outline the features of the theory that are relevant for the present work, in particular we recall the postulates regarding the structure of state spaces and the characterization of channels. For a detailed account we refer the reader to [44].

Postulate 4. *Bilocal Classical Theory is classical and convex. For every integer $D > 1$ there exists a type of system with size D.*

Postulate 5 (Parallel composition)**.** *For any two systems* A, B*, the size of the composite system* AB *is given by the following rule:*

$$D_{\mathrm{AB}} = D_{\mathrm{BA}} = \begin{cases} 2D_{\mathrm{A}}D_{\mathrm{B}}, & \text{if } \mathrm{A} \neq \mathrm{I} \neq \mathrm{B}, \\ D_{\mathrm{A}}, & \text{if } \mathrm{B} = \mathrm{I}. \end{cases} \tag{3.42}$$

Let $\mathrm{I} \neq \mathrm{A}, \mathrm{B}, \mathrm{C}$*. We denote the set of pure states of any composite system* AB *as* $\mathsf{PurSt}(\mathrm{AB}) = \{|(ij)_-)_{\mathrm{AB}}, |(ij)_+)_{\mathrm{AB}} \mid 1 \leq i \leq D_{\mathrm{A}}, 1 \leq j \leq D_{\mathrm{B}}\}$*, so that for all pure states* $|i)_{\mathrm{A}} \in \mathsf{PurSt}(\mathrm{A}), |j)_{\mathrm{B}} \in \mathsf{PurSt}(\mathrm{B})$ *the following parallel composition rule holds:*



(3.43)

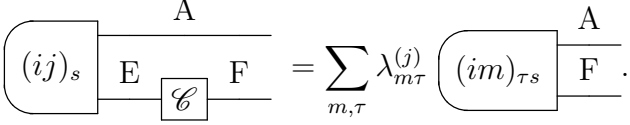
When a third system is added the association satisfies the following law:

$$((ij)_{s_1}k)_{s_2} = (i(jk)_{s_1 s_2})_{s_1}, \tag{3.44}$$

for all local indices i, j, k and signs s_1, s_2.

The following proposition is actually a consequence of other defining postulates that we have not mentioned here (the interested reader can find all details in the original reference [44]). It tells us how channels are characterised, and it is of fundamental importance when we will discuss the coding problem in this theory (which is connected to the information content, as we will describe in-depth in chapter 4), since it essentially defines how compression schemes can be constructed in this theory

Property 4 (Channels)**.** $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A} \to \mathrm{B})$ *iff for every* $i \in \{1, \ldots, D_{\mathrm{A}}\}$ *there exists a probability distribution* $\{\lambda^{(i)}_{m,\tau}\}_{(m,\tau)\in I}$*, with* $I = \{1, \ldots, D_{\mathrm{B}}\} \times \{+, -\}$*, such that the following holds for all* $|(ij)_s)_{\mathrm{AE}} \in \mathsf{PurSt}(\mathrm{AE})$*:*



3.3 Fidelity and entropic-like quantities in OPT

Several information-theoretic quantities find their place in this framework and all the ones that are summarized in this section rely on their classical counterparts. For instance a notion of fidelity can be suitably defined based on the classical fidelity[4] (or Bhattacharyya coefficient) [54, 75]. Moreover in the literature, three notions of entropies have been introduced and studied for GPTs and OPTs [54, 55, 56, 76, 77, 78], and each of them has its own operational interpretation that is directly inherited from the intuitive understanding of the Shannon entropy as a measure of uncertainty. While it is known that they all collapse to the same quantity in CT and QT, the Shannon and von Neumann entropies respectively, it has also been noticed that they are not equivalent for generalised probabilistic models. In this section we recall and discuss their definitions and we remind the reader of some of the relevant properties they satisfy.

3.3.1 Fidelity

The fidelity is an alternative way of quantifying the distance between states that is widely used in information theory. In QT, we have seen that the fidelity F between two density matrices ρ and σ is defined via the formula $F(\rho,\sigma) = \mathrm{Tr}(\sqrt{\rho^{1/2}\sigma\rho^{1/2}})$, but can also be obtained as the minimization over all the possible POVMs of the classical fidelity, as expressed in equation (2.29). By reversing the logic, one can use this characterization of F that holds true for QT and use it as a definition in the context of OPT.

Definition 3.3.1. *Let $\rho,\sigma \in \mathsf{St}_1(\mathrm{A})$. For any observation test $\{a_i\}_{i\in\mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{A})$ denote by $\mathbf{p} \coloneqq p_i$ and $\mathbf{q} \coloneqq q_i$ the probability distributions defined by*

$$p_i \coloneqq \rho \xrightarrow{\mathrm{A}} a_i \,, \qquad q_i \coloneqq \sigma \xrightarrow{\mathrm{A}} a_i \,.$$

Then one can define the Fidelity between ρ and σ as

$$F(\rho,\sigma) \coloneqq \inf_{\{a_i\}_{i\in\mathsf{X}}} F_c(\mathbf{p},\mathbf{q}), \tag{3.45}$$

where $F_c(\mathbf{p},\mathbf{q}) = \sum_i \sqrt{p_i q_i}$.

Since the classical fidelity is bounded by 1, and is equal to 1 only for $\mathbf{p} = \mathbf{q}$, one clearly has $0 \leq F(\rho,\sigma) \leq 1$, with equality if and only if $\rho = \sigma$; moreover, it is symmetric in its arguments, namely $F(\rho,\sigma) = F(\sigma,\rho)$. Fidelity satisfies the following property, that generalises the Fuchs-van de Graaf inequality [30] and establishes the equivalence with the operational distance as a measure of closeness on the set of deterministic states.

[4] Given two probability distributions $\mathbf{p} \coloneqq \{p_i\}_{i\in\mathsf{X}}$ and $\mathbf{q} \coloneqq \{q_i\}_{i\in\mathsf{X}}$ one can define the classical fidelity as $F_c(\mathbf{p},\mathbf{q}) = \sum_i \sqrt{p_i q_i}$.

Proposition 3.3.1. *Let $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$. The following inequalities hold*

$$1 - F(\rho,\sigma) \leq \frac{1}{2}\|\rho - \sigma\|_{\mathrm{op}} \leq \sqrt{1 - F(\rho,\sigma)^2}. \tag{3.46}$$

Proposition 3.3.2 (Monotonicity)**.** *Let $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A},\mathrm{B})$. The following inequality holds*

$$F(\mathscr{C}(\rho), \mathscr{C}(\sigma)) \geq F(\rho,\sigma). \tag{3.47}$$

Beyond these two properties, that have their own statement just for future reference, we notice that this generalized fidelity is also strongly concave (hence concave and jointly concave) [54].

3.3.2 Measurement entropy

A characterisation of the Shannon entropy in QT can be given in terms of a minimization procedure of the Shannon entropy. In more precise terms, take an arbitrary rank-one POVM $\{\Lambda_i\}_{i\in\mathsf{X}}$ and, for a given state ρ, consider the Shannon entropy of the probability distribution $\mathbf{p} := \{\mathrm{Tr}(\rho\Lambda_i)\}$. Then, it can be shown that the von Neumann entropy of the state ρ can be computed by taking the infimum over all such POVMs, namely $S(\rho) := \inf_{\Lambda_i} H(\mathbf{p})$. Actually, one can replace the infimum with the minimum since $S(\rho)$ can be achieved by a projective measurement associated with the eigenstates of the quantum state ρ.

As for fidelity, one can consider this as a definition of entropy. Hence, given any state $\rho \in \mathsf{St}_1(\mathrm{A})$, one can consider all the possible atomic measurements $\{a_i\}_{i\in\ \mathsf{X}}$ that can be performed thereon, each one defining a probability distribution $\mathbf{p} = \{(a_i|\rho)\}_{i\in\mathsf{X}}$. Therefore, the uncertainty associated with this measurement is obtained by computing the Shannon entropy of $\mathbf{p}$, and then optimizing by taking the infimum.

Definition 3.3.2. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ for some system A, denote by $\mathcal{O}^{\mathrm{at}}(\mathrm{A})$ the set of atomic observation test of A. We then define*

$$S_1(\rho) := \inf_{\{a_i\}\in\mathcal{O}^{\mathrm{at}}(\mathrm{A})} H(a_i(\rho).) \tag{3.48}$$

Now the infimum is mandatory, since it could be the case that an observation test that achieves it does not exist. Notice that an optimization over all the set of masurements would give a trivial result, since one can always choose the measurement corresponding to a deterministic effect $\{e_{\mathrm{A}}\}$, so that $S_1(\rho) = 0$; as it is clear from the definition, this can be interpreted as the least uncertainty among all the possible atomic measurements. Concerning its properties, $S_1(\rho)$ is a non-negative concave function, and if the parallel composition of atomic measurements yields a measurement which is still atomic, then it is also subadditive. However, it is also shown [56] that there exists a tripartite state of non-local boxes for which strong subadditivity (see item 6 of theorem 2.1.1) is violated.

Remark 3.3.1. In [56] the authors considered a compression task and analyse, under certain additional assumptions, the connection of the compression rate with this definition entropy. In the first place, for a subset of S of a set of states $\mathsf{St}(\mathrm{A})$, they define a notion of dimension as $d = \min_{\{a_i\}_{i\in\mathsf{X}}\in\mathcal{O}^{\mathrm{at}}} |\{i \in \mathsf{X} | \exists \sigma \in S, (a_i|\sigma) > 0\}|$. In other words, d is computed by considering an atomic measurement $\{a_i\}_{i\in\mathsf{X}}$, then counting the number of outcomes in X such that at least one state in S that is different from zero, and eventually minimising this number over all the atomic measurements. Then they define an i.i.d. source as an ensemble $\{q_i, \sigma_i\}_{i\in\mathsf{X}}$ which emits sequences of states of the form $\sigma_{\mathbf{i}} := \sigma_{i_1} \boxtimes \cdots \boxtimes \sigma_{i_N}$. These states lie in the product space $\mathsf{St}(\mathrm{A}^{\otimes N})$ and the encoding maps these to states σ^N in a subset $S^N \subseteq \mathsf{St}(A^{\otimes N})$. After the decoding we end up with a state $\tilde{\sigma}_{\mathbf{i}}$, and the compression scheme has rate R if the dimension of S^N satisfies $d_{S^N} \leq 2^{NR}$. The reliability is assessed by means of the average error distance between the original state and the decoded one, $\sum_{\mathbf{i}} q_{\mathbf{i}} \|\sigma_{\mathbf{i}} - \tilde{\sigma}_{\mathbf{i}}\|_{\mathrm{op}}$. Thus, setting $\rho = \sum_i q_i \sigma_i$, they prove that if $R > S_1(\rho)$ then the source can be reliably compressed, namely that $S_1(\rho)$ is an achievable rate for this task.

3.3.3 Accessible information

Given any $\rho \in \mathsf{St}(\mathrm{A})$, for any convex decomposition $\{p_i, \sigma_i\}_{i\in\mathsf{X}}$ one can define the accessible information associated with it by optimising over all the possible measurements that can be performed thereon as

$$I(\{p_i, \sigma_i\}) := \sup_{\{a_j\}_{j\in\mathsf{Y}}} H(X : J),$$

where the mutual information is computed via the joint probability distribution defined as $p(i, j) := p_i(a_j|\sigma_i)_{\mathrm{A}}$. Then, one can define an entropy by optmising over all the possible convex decompositions of ρ. However, one obtains the same result if pure convex decompositions and atomic measurements are considered, so that one can equivalently take the following as a definition.

Definition 3.3.3. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ for some system A, denote by $\mathcal{O}^{\mathrm{at}}(\mathrm{A})$ the set of atomic observation test of A. We then define*

$$S_2(\rho) := \sup_{(p_i,\{\phi_i\})\in\mathcal{P}(\rho)} \sup_{\{a_i\}\in\mathcal{O}^{\mathrm{at}}(\mathrm{A})} H(X : J) \tag{3.49}$$

where the mutual information is computed according to the joint probability distribution $p_i(a_j|\phi_i)_{\mathrm{A}}$.

Non-negativity of S_2 directly follows by non negativity of the Shannon mutual information. An important feature of this entropy is that it can be understood as a measure of pureness, in the sense that $S_2(\rho)$ is vanishing if and only if ρ is a pure state.

Remark 3.3.2. Contrary to the measurement entropy, S_2 is generally not concave but, if $\{\sigma_i\}_{i\in\mathsf{X}}$ is a collection of states and $\mathbf{p} = \{p_i\}_{i\in\mathsf{X}}$ is a probability distribution, then the following inequalities hold in any probabilistic model

$$\begin{aligned}
S_2(\sum_i p_i\sigma_i) &\geq \frac{\sum_i p_i^2 S_2^2(\sigma_i)}{\sum_i p_i S_2(\sigma_i)},\\
S_2(\sum_i p_i\sigma_i) &\geq \sum_i p_i^2 S_2(\sigma_i),\\
S_2(\sum_i p_i\sigma_i) &\geq \frac{1}{|\mathsf{X}|}\sum_i p_i S_2(\sigma_i),\\
S_2(\sum_i p_i\sigma_i) &\geq \max_{i\in\mathsf{X}} p_i S_2(\sigma_i).
\end{aligned}$$

These relations express a weaker form of concavity, since they are a straightfoward consequence of the latter.

3.3.4 Mixing entropy

Let $\rho \in \mathsf{St}(\mathrm{A})$. This state may be decomposed in terms of pure states in several ways. If $\rho = \sum_i p_i\phi_i$ with $\mathbf{p} = \{p_i\}_{i\in\mathsf{X}}$ a probability distribution and $\{\phi_i\}$ a set of pure states, then one can compute Shannon entropy $H(\mathbf{p})$ associated with such a pure decomposition, which measures the uncertainty associated with it. By minimising over all the possible decompositions we are led to the following definition.

Definition 3.3.4. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ for some system* A*. We then define*

$$S_3(\rho) := \inf_{\{(p_i,\phi_i\})\in\mathcal{P}(\rho)} H(p_i). \tag{3.50}$$

As for S_2, also the mixing entropy is vanishing on all and only the pure states [54]. $S_3(\rho)$ measures the least uncertainty over all the possible way of decomposing ρ in pure states and it exhibits the weirdest behaviour. For instance, it is proven to be not subadditive, and when the set of states is not a simplicial polytope it is not concave. However, if $\{\sigma_i\}_{i\in\mathsf{X}}$ is a collection of states and $\mathbf{p} = \{p_i\}_{i\in\mathsf{X}}$ is a probability distribution, then it holds that

$$S_3(\sum_i p_i\sigma_i) \leq H(\mathbf{p}) + \sum_i p_i S_3(\sigma_i),$$

which is an upper bound that is also satisfied by the von Neumann entropy.

As mentioned at the beginning of this section, there are probabilistic models where the mixing entropy differs from the ones we have already described. In [55] the authors provide a definition of mutual information in terms of the measurement entropy, which is just the same definition of classical information theory, where $S_1(\rho)$ replaces the Shannon entropy, and they provide a sufficient condition for information causality based on this mutual information. In

particular, they show that a theory where i) the mixing entropy is equal to the measurement entropy, ii) the mutual information satisfies a Holevo bound and iii) the measurement entropy is strongly subadditive, then satisfies the information causality principle introduced in [79].

CHAPTER 4

The information content in OPT

In this chapter we introduce a notion of information content for the OPT framework, based on a compression task that generalises both the quantum and the classical setting that we revised in chapter 2. The requirement that we impose on compression schemes featuring in the definition of information content is that their effect on any preparation of ensembles that average to the considered state must be indistinguishable from leaving the preparation untouched. Thus, besides considering any refinement of the state under discussion, we consider the action of the compression scheme on decompositions of its *dilations*, i.e. joint states of the system and an arbitrary external system such that the state that one obtains after averaging and discarding the external system is precisely the one of interest. The importance of considering the effect of transformations on external systems has been already pointed out in other contexts, e.g. in assessing information vs disturbance [52].

In classical and quantum Shannon theories the amount of information is measured in bits and qubits respectively. For a generic theory, with no further restrictions on its structure, in principle, one may not be able to identify an elementary information carrier. For this reason, we will consider theories that we name "digitisable". Roughly speaking, we assume the existence of at least one elementary system, which we call obit[1], such that an agent can always encode an arbitrary but finite number of copies of her/his system into an array made of an integer number of obits. This feature is evidently satisfied by classical and quantum theory, and it does not rule out scenarios that are relevant

[1]Here, "o" stands for operational.

from a foundational perspective, such as non-local boxes [80, 81]. The latter is the prototypical example of a theory where the various notions of entropy exhibit odd features, such as violation of strong subadditivity, subadditivity and concavity, and where they are also proven to be not equivalent [55, 56, 82]. Moreover, this assumption can also be applied to theories without local tomography, fermionic information theory (that we will analyse later) and BCT, that will be discussed in chapter 6.

After introducing the notion of information content in a general OPT, we prove some of its main properties and, as special cases, we analyse classical and quantum theory, where our definition boils down to Shannon's and von Neumann's entropy, respectively. We then analyse in detail when the information content can be understood as a measure of purity for states and, as a byproduct, we also show that the optimised accessible information [55, 56, 82] generally provides a lower bound for it.

4.1 Definition of information content

In this section we present an operational definition of information content, which is essentially based on a generalisation of the compression task reviewed in detail in chapter 2. We first give the notion of digitisable OPT, that in turn set the stage for the introduction of the information content, which we define immediately after. The main ingredient of the definition is the figure of merit, that establishes how much the original message emitted by the source is distorted by compression protocols. We then show that in theories that satisfy conditioning and steering (properties 2 and 3) the figure of merit can be computed according to simplified expressions, bringing them closer to the error quantifiers already introduced in the classical and quantum settings. We then show that in the latter cases the information content reduces to the Shannon and von Neumann entropy respectively.

4.1.1 Digitisable theories

In a general physical theory of information, information plays of course a central role. It is then reasonable to expect that one has a way to quantify it, and this in turn has to come along with a unit. In classical and quantum theories, the output of any physical source is digitised in terms of bits and qubits, respectively. In particular, focusing on the case of classical information theory, any string of length N, say $\mathbf{i} := i_1 \dots i_N$, where each i is extracted from an alphabet of d symbols, can be perfectly encoded on a distinct array of a suitable number of bits. The bit is indeed the reference system that we usually adopt as a unit for assessing the amount of information of a given classical source; an analogous role is played by the qubit in quantum information theory. Moreover, notice that one can actually choose any other system for the digitisation of the outputs of a classical or quantum information source, with

the only difference being that the entropy function, which numerically quantifies the information content, must be rescaled by a suitable multiplicative factor.

In order to elaborate more on this, let us focus on QT. Consider two different qudits with dimension d_1 and d_2 respectively. Generally, the equation $d_1^{k_1} = d_2^{k_2}$ may have no integer solutions (for instance, when both d_1 and d_2 are prime). However, such equation tells us that the smallest integer $k_2 := M^{\min}(k_1)$ such that we can isometrically embed $\mathcal{H}_1^{\otimes k_1}$ into $\mathcal{H}_2^{\otimes k_2}$ is trivially given by

$$M^{\min}(k_1) = \lceil k_1 \log_{d_2} d_1 \rceil, \tag{4.1}$$

which is such that

$$\lim_{k_1 \to \infty} \frac{\lceil k_1 \log_{d_2} d_1 \rceil}{k_1} = \log_{d_2} d_1. \tag{4.2}$$

Similarly, $\lceil k_2 \log_{d_1} d_2 \rceil$ is needed for an isometric embedding of $\mathcal{H}_2^{\otimes k_2}$ into $\mathcal{H}_1^{\otimes k_1}$, and one has that

$$\lim_{k_2 \to \infty} \frac{\lceil k_2 \log_{d_1} d_2 \rceil}{k_2} = \log_{d_1} d_2 = \frac{1}{\log_{d_2} d_1} = \frac{1}{\lim_{k_1 \to \infty} \frac{\lceil k_1 \log_{d_2} d_1 \rceil}{k_1}}. \tag{4.3}$$

This simple observation tells us that any pair of qudit systems A_1 and A_2 are *asymptotically equivalent systems* according to the following definition.

Definition 4.1.1 (Asymptotical equivalence)**.** *Two systems* A_1 *and* A_2 *are asymptotically equivalent if*

1. *there exists a pair of integers* $k_1, k_2 < \infty$, $\mathscr{E} \in \mathsf{Transf}_1(\mathrm{A}_1^{\boxtimes k_1} \to \mathrm{A}_2^{\boxtimes k_2})$ *and* $\mathscr{D} \in \mathsf{Transf}_1(\mathrm{A}_2^{\boxtimes k_2} \to \mathrm{A}_1^{\boxtimes k_1})$ *such that* $\mathscr{D}\mathscr{E} = \mathscr{I}_{\mathrm{A}_1^{\boxtimes k_1}}$*;*

2. *there exists a pair of integers* $h_1, h_2 < \infty$, $\mathscr{G} \in \mathsf{Transf}_1(\mathrm{A}_2^{\boxtimes h_2} \to \mathrm{A}_1^{\boxtimes h_1})$ *and* $\mathscr{F} \in \mathsf{Transf}_1(\mathrm{A}_1^{\boxtimes h_1} \to \mathrm{A}_2^{\boxtimes h_2})$ *such that* $\mathscr{F}\mathscr{G} = \mathscr{I}_{\mathrm{A}_2^{\boxtimes h_2}}$*;*

3. *let* $M_2^{\min}(k_1)$ *be the smallest* k_2 *such that item 1 is satisfied for a given* k_1*, and similarly for* $M_1^{\min}(h_2)$ *with reference to item 2. The following assumption is made:*

$$\lim_{k_1 \to \infty} \frac{M_2^{\min}(k_1)}{k_1} = k, \qquad \lim_{h_2 \to \infty} \frac{M_1^{\min}(h_2)}{h_2} = k^{-1}. \tag{4.4}$$

We can summarise the observations that we have made up until now as follows:

1. given any quantum system A, associated with an Hilbert space $\mathcal{H}_{\mathrm{A}}$, there exists at least an elementary system Q and an integer M such that there exists a left-reversible quantum channel $\mathscr{E} : \mathrm{A} \to \mathrm{Q}^{\otimes M}$;

2. if Q_1 and Q_2 are two elementary systems, then they are asimptotically equivalent.

These two simple considerations can be thoroughly made into a requirement that a theory must abide by, thus setting the stage for a meaningful definition of compression rate in the OPT framework. This is formalized as follows.

Assumption 1 (Digitisability)**.** *We say that an OPT is digitizable if there exists a system* B *(called* obit*) such that for any system* X *there exists* $k < \infty$ *and a pair of maps* $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{X} \to \mathrm{B}^{\boxtimes k})$ *and* $\mathscr{F} \in \mathsf{Transf}_1(\mathrm{B}^{\boxtimes k} \to \mathrm{X})$ *such that* $\mathscr{F} \circ \mathscr{C} = \mathscr{I}_{\mathrm{X}}$*. Moreover, if* B_1 *and* B_2 *are two such systems, then they are asymptotically equivalent.*

In any theory satisfying the assumption of digitisability an agent can always encode the state of her/his system on the parallel composition of a sufficiently large number of elementary systems, which we can think of as a generalisation of the qubit system for quantum theory. The request of digitisability comes from the need of a unit for the amount of information required in order to store a given source. In classical information theory we use bits, in the quantum counterpart the qubits, and for a generic OPT satisfying digitizability we use obits, whose existence must then be postulated.

Remark 4.1.1. We want to stress the fact that the assumption of digitisability is extremely weak, to the extent that every theory in the literature abides by it, and it is very hard to imagine a theory that violates it. Indeed, a non-digitizable theory should contain infinitely many inequivalent system types, even asymptotically (see def. 4.1.1), and this immediately brings us into an unexplored territory of wild theories.

Remark 4.1.2. The non-local boxes [80] provide us with a non trivial example of an OPT satisfying the digitisability assumption, and with a strong departure from the quantum one. There exists a unique single system, whose state space is described by a square, and multipartite systems are obtained by using only this one, therefore it is trivially digitizable. In the literature—with a few remarkable exceptions [41, 83]—non-local boxes are presented in terms of the geometry of their state space, and focusing on the correlations that measurements can produce, disregarding the behaviour of transformations. However one can straightforwardly make them into an OPT by assuming that every collection of linear maps on the state space that map preparations to preparations is allowed. A similar construction was carried out e.g. in Ref. [41]. Non-local boxes provide a scenario where the wealth of the known entropy notions is manifest. Moreover, the fact that we are only referring to the conversion of finitely many copies of the system at hand is not constraining from a conceptual point of view. For the present purposes, namely taking a first step towards a Shannon theory for generic physical systems, this level of analysis is sufficient. However, the composition of a countable number of systems can be suitably defined (see [63]) opening the route to a generalization of this property in the infinite case.

4.1.2 The definition

We now introduce the definition of information content in terms of a compression task, where the procedure is defined in terms of a block-coding strategy where we allow for an error at the decoding, mimicking the scenario that we have described at length in chapter 2. An *information source* is characterized by a system A and a state $\rho \in \mathsf{St}(\mathrm{A})$. Repeated uses of the source then generate a *message* that, generalizing the i.i.d. setting of the classical and quantum scenario, has the form of a factorized state, $\rho^{\boxtimes N}$. A compression map is given by a pair of maps $\mathscr{E} \in \mathsf{Transf}_1(\mathrm{A}^{\boxtimes N} \to \mathrm{B}^{\boxtimes M})$ and $\mathscr{D} \in \mathsf{Transf}_1(\mathrm{B}^{\boxtimes M} \to \mathrm{A}^{\boxtimes N})$, that we name *encoding* and *decoding* respectively; we also call $\mathscr{C} := \mathscr{D}\mathscr{E}$ the *codec* map.

Now, we need a figure of merit that establishes the goodness of the codec map $\mathscr{C} := \mathscr{D}\mathscr{E}$ or, in other words, we must define a quantity that gauges how much error the codec map $\mathscr{C}$ introduces. Our choice comes as a consequence of the two following elementary observations: (i) the output message to which we have local access may be correlated with an ancilla, and (ii) it could be prepared in different ways. These considerations come from classical and quantum Shannon theory, which taught us that faithfully transmitting information amounts to distorting both preparations and correlations as little as possible. Therefore, since the aim is to preserve all the possible information gathered in the message, we use the following quantity

$$D(\rho^{\boxtimes N}, \mathscr{C}) := \sup_{\Psi \in D_{\rho^{\boxtimes N}}} \sup_{\{\Gamma_i\} \in \mathcal{R}(\Psi)} \sum_i \|\mathscr{C} \boxtimes \mathscr{I}(\Gamma_i) - \Gamma_i\|_{\mathrm{op}}, \tag{4.5}$$

as the error quantifier. In other words, it is an optimization of the average error over all the possible decompositions of any dilation of the message.

Accordingly, we can now give the following definitions

Definition 4.1.2. *An* (ε, N)-reliable compression scheme $(\mathscr{E}, \mathscr{D})$ *is a compression scheme such that* $D(\rho^{\boxtimes N}, \mathscr{C}) < \varepsilon$ *where* $D(\rho^{\boxtimes N}, \mathscr{C})$ *is computed according to formula* (4.5). *For fixed* N, M *we denote with* $E_{N,M,\varepsilon}(\rho)$ *the set of* ε*-reliable compression schemes, i.e.*

$$E_{N,M,\varepsilon}(\rho) = \{(\mathscr{E}, \mathscr{D}) | D(\rho^{\boxtimes N}, \mathscr{C}) < \varepsilon\}. \tag{4.6}$$

Definition 4.1.3 (Information Content)**.** *Let* $\rho \in \mathsf{St}_1(\mathrm{A})$. *We define the* smallest achievable compression rate *for length* N *to tolerance* ε *as follows*

$$R_{N,\varepsilon}(\rho) \coloneqq \frac{\min\{M : E_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N}. \tag{4.7}$$

The information content *of the state* ρ *is defined as*

$$I(\rho) \coloneqq \lim_{\varepsilon \to 0} \limsup_{N \to \infty} R_{N,\varepsilon}(\rho). \tag{4.8}$$

In the following proposition we show that in any digitisable theory the information content in well-defined.

Proposition 4.1.1. *$I(\rho)$ is well defined for every $\rho \in \mathsf{St}_1(\mathrm{A})$ and every system* A*.*

Proof. Firstly, we show that for any choice of the elementary system, $I(\rho)$ is a finite number for any state ρ. By the digitizability assumption we know that for any N there exists a positive integer $K < \infty$ and a pair of maps $\mathscr{E} \in \mathsf{Transf}_1(\mathrm{A}^{\boxtimes N} \to \mathrm{B}^{\boxtimes K})$, $\mathscr{D} \in \mathsf{Transf}_1(\mathrm{B}^{\boxtimes K} \to \mathrm{A}^{\boxtimes N})$ such that $\mathscr{D}\mathscr{E} = \mathscr{I}_{\mathrm{A}^{\boxtimes N}}$. Therefore, for any N, ε the set $E_{N,K,\varepsilon}(\rho)$ is not empty, and the minimum in eq. (4.7) is always finite. Moreover, it is immediate to realise that K does not need to grow more than linearly versus N, just considering N repetitions of the encoding for one copy A. Thus, the ratio in eq. (4.7) is bounded and one can take the $\limsup_{N\to\infty}$ safely. The existence of $\lim_{\varepsilon\to 0}$ follows by the fact that $E_{N,M,\varepsilon}(\rho) \subseteq E_{N,M,\varepsilon'}(\rho) \neq \emptyset$ whenever $\varepsilon \leq \varepsilon'$, which, in turn, implies monotonicity of the function $\limsup_{N\to\infty} R_{N,\varepsilon}(\rho)$ versus ε.

What is left to prove is that using two different obits we are not led to two incomparable notions of information content. First of all, fix N, ε, let $\rho \in \mathsf{St}_1(\mathrm{A})$, and let

$$M_{1,N} \coloneqq \min\{M : E^1_{N,M,\varepsilon}(\rho) \neq \emptyset\},$$
$$M_{2,N} \coloneqq \min\{M : E^2_{N,M,\varepsilon}(\rho) \neq \emptyset\},$$

be the minimum number of obits B_1 and B_2 needed for an ε-optimal encoding, respectively. Rephrasing the first equation in (4.4), there exists a sequence $\delta_1(M_1)$ such that $M_2^{\min}(M_1) = kM_1 + \delta_1(M_1)$ with $\lim_{M_1\to\infty} \frac{\delta_1(M_1)}{M_1} = 0$. Given the encoding $\mathscr{E}$ of item 1 in definition 4.1.1 from $M_{1,N}$ to $M_2^{\min}(M_{1,N})$ (see item 3 in definition 4.1.1), we have an ε-optimal encoding of $\rho^{\boxtimes N}$ onto $M_2^{\min}(M_{1,N})$ obits B_2, therefore $M_{2,N} \leq M_2^{\min}(M_{1,N})$ and this implies

$$\begin{aligned}
\limsup_{N\to\infty} \frac{M_{2,N}}{N} &\leq \limsup_{N\to\infty} \frac{M_2^{\min}(M_{1,N})}{N} \\
&= \limsup_{N\to\infty} \left[\frac{kM_{1,N}}{N} + \frac{\delta_1(M_{1,N})}{N}\right] \\
&\leq \limsup_{N\to\infty} \frac{kM_{1,N}}{N} + \limsup_{N\to\infty} \left|\frac{\delta_1(M_{1,N})}{M_{1,N}} \frac{M_{1,N}}{N}\right| \\
&= k \limsup_{N\to\infty} \frac{M_{1,N}}{N}.
\end{aligned}$$

The last line follows by $\lim_{M_1\to\infty} \frac{\delta_1(M_1)}{M_1} = 0$ along with the fact that $M_{1,N+1} \geq M_{1,N}$. Taking $\lim_{\varepsilon\to 0}$ we end up with $I_2(\rho) \leq kI_1(\rho)$. A similar argument can be used to show the reverse inequality, and we have that, for any $\rho \in \mathsf{St}_1(\mathrm{A})$, $I_2(\rho) = kI_1(\rho)$. □

4.1.3 Alternative figure of merits

As it is clear, the definition of $I(\rho)$ strongly depends on the choice of the figure of merit $D(\rho^{\boxtimes N}, \mathscr{C})$. The one that we have introduced takes into account all

possible refinements of any dilation of $\rho^{\boxtimes N}$. Alternatively, one can consider the pure convex refinements of any dilation of $\rho^{\boxtimes N}$ and define

$$D^{\mathrm{pur}}(\rho^{\boxtimes N}, \mathscr{C}) := \sup_{\Psi \in D_{\rho^{\boxtimes N}}} \sup_{\{p_i, \Psi_i\} \in \mathcal{P}(\Psi)} \sum_i p_i \|\mathscr{C} \boxtimes \mathscr{I}(\Psi_i) - \Psi_i\|_{\mathrm{op}}. \tag{4.9}$$

When the theory satisfies strong causality (property 2), and if any state is proportional to a determistic one, D and D^{pur} can be equivalently used to define the information content via the formula (4.8).

Proposition 4.1.2. *Consider an OPT satisfying strong causality and where every state $\sigma \in \mathsf{St}(\mathrm{B})$ is proportional to a deterministic one for every system* B*. Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and consider*

$$I^{\mathrm{pur}}(\rho) \coloneqq \lim_{\varepsilon \to 0} \limsup_{N \to \infty} \frac{\min\{M : E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N}, \tag{4.10}$$

where $E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho)$ is the following set

$$E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho) = \{(\mathscr{E}, \mathscr{D}) | D^{\mathrm{dil}}(\rho^{\boxtimes N}, \mathscr{C}) < \varepsilon\}. \tag{4.11}$$

Then $I(\rho) = I^{\mathrm{pur}}(\rho)$.

Proof. On the one hand, we trivially have $I^{\mathrm{pur}}(\rho) \leq I(\rho)$. On the other hand, let $\{\Psi_i\}_{i \in \mathsf{X}}$ be a refinement of $\Omega \in D_{\rho^{\boxtimes N}}$. Any element in the refinement is proportional to a deterministic state, so that there exists $\{\Gamma_i\}_{i \in \mathsf{X}} \subseteq \mathsf{St}_1(\mathrm{A}^{\boxtimes N}\mathrm{E})$ and a probability distributions such that $\{(p_i, \Gamma_i)\}_{i \in \mathsf{X}}$ is a convex refinement. Conditioning implies that for every Γ_i there exists a convex refinement made of pure states, namely there exists a probability distribution $\{q_{j|i}\}_{j \in \mathsf{Y}_i}$ and a set of pure states $\{\Phi^{(i)}_j\}_{j \in \mathsf{Y}_i}$ such that $\{q_{j|i}\Phi^{(i)}_j\}_{j \in \mathsf{Y}_i}$ is a refinement of Γ_i. Moreover, conditioning also implies that $\cup_{i \in \mathsf{X}}\{(p_i, q_{j|i}\Phi^{(i)}_j)\}_{j \in \mathsf{Y}_i}$ is a pure convex refinement of Ω. Thus, by the triangle inequality one has

$$\begin{aligned}
&\sum_{i \in \mathsf{X}} \|[(\mathscr{D} \circ \mathscr{C}) \boxtimes \mathscr{I}]\Psi_i - \Psi_i\|_{\mathrm{op}} \\
&\leq \sum_{i \in \mathsf{X}} \sum_{j \in \mathsf{Y}_i} p_i q_{j|i} \|[(\mathscr{D} \circ \mathscr{C}) \boxtimes \mathscr{I}]\Phi^{(i)}_j - \Phi^{(i)}_j\|_{\mathrm{op}}.
\end{aligned}$$

This implies that $E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho) \subseteq E_{N,M,\varepsilon}(\rho)$, and in turn that $I(\rho) \leq I^{\mathrm{pur}}(\rho)$ Therefore $I(\rho) = I^{\mathrm{pur}}(\rho)$. □

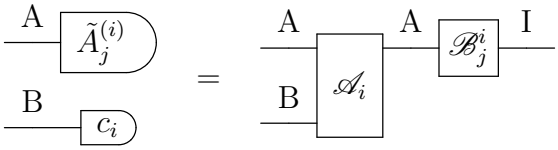
If a theory also satisfies steering (property 3), along with strong causality, one can use a figure of merit that is further simplified

$$D^{\mathrm{dil}}(\rho^{\boxtimes N}, \mathscr{C}) := \sup_{\Psi \in D_{\rho^{\boxtimes N}}} \|\mathscr{C} \boxtimes \mathscr{I}(\Psi) - \Psi\|_{\mathrm{op}}. \tag{4.12}$$

This is proved in proposition 4.1.3, but we need the following two lemmas first.

Lemma 4.1.1. *Let* $\{c_i\}_{i\in\mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{B})$ *be an observation test and* $\{A\}_{i\in\mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{A})$ *a collection of effects. If causality holds, then* $\sum_{i\in\mathsf{X}} A_i \boxtimes c_i \in \mathsf{Eff}(\mathrm{AB})$.

Proof. This is a straightforward consequence of conditioning. For any $i \in \mathsf{X}$, there exists an observation test $\{\tilde{A}_j^{(i)}\}_{j\in\mathsf{Y}_i}$ such that $A_i \in \{\tilde{A}_j^{(i)}\}_{j\in\mathsf{Y}_i}$. Thus, if we consider the test $\{\mathscr{I}_{\mathrm{B}}\boxtimes c_i\}_{i\in\mathsf{X}}$ and the collection of effects $\{\tilde{A}_j^{(i)}\boxtimes c_i\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$ we have



(4.13)

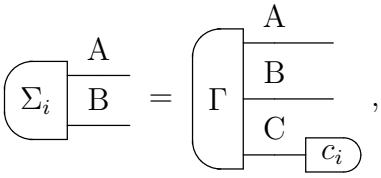
with $\mathscr{A}_i \coloneqq \mathscr{I}_{\mathrm{A}} \boxtimes c_i$ and $\mathscr{B}_j^i \coloneqq \tilde{A}_j^{(i)}$. Therefore, strong causality implies that $\{\tilde{A}_j^{(i)} \boxtimes c_i\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$ is an observation test, and $\sum_{i\in\mathsf{X}} A_i \boxtimes c_i \in \mathsf{Eff}(\mathrm{AB})$, being a coarse graining of effects from the same test. □

Lemma 4.1.2. *Let* $\rho \in \mathsf{St}_1(\mathrm{A})$, $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A})$ *and* $\varepsilon > 0$. *If* $\|\mathscr{C}\boxtimes\mathscr{I}(\Psi)-\Psi\| < \varepsilon$ *for any* $\Psi \in D_\rho$ *and assumption 3 holds, then one has*

$$\sum_{i\in\mathsf{X}} \|\mathscr{C} \boxtimes \mathscr{I}(\Sigma_i) - \Sigma_i\|_{\mathrm{op}} < \varepsilon,$$

for any refinement $\{\Sigma_i\}_{i\in\mathsf{X}}$ *of an element of* D_ρ.

Proof. Let $\{\Sigma_i\}_{i\in\mathsf{X}}$ be the refinement of an element Ω of D_ρ. By Property 3 there exists $\Gamma \in D_\Omega \subseteq D_\rho$ and an observation test $\{c_i\}_{i\in\mathsf{X}}$ such that



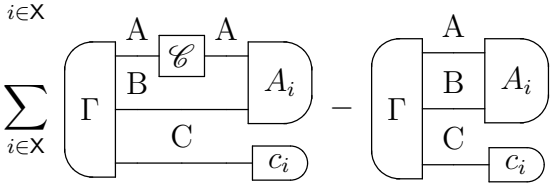
, $\qquad \forall i \in \mathsf{X}.$

For any $i \in \mathsf{X}$, let $A_i \in \mathsf{Eff}(\mathrm{AB})$ be the effect achieving the norm, namely such that

$$\|\mathscr{C} \boxtimes \mathscr{I}(\Sigma_i) - \Sigma_i\|_{\mathrm{op}} = (A_i|[(\mathscr{C} - \mathscr{I}) \boxtimes \mathscr{I}]|\Sigma_i).$$

Since $\sum_{i\in\mathsf{X}} A_i \boxtimes c_i$ is an effect (by lemma 4.1.1), we have that

$$\sum_{i\in\mathsf{X}} \|\mathscr{C} \boxtimes \mathscr{I}(\Sigma_i) - \Sigma_i\|_{\mathrm{op}} =$$



$$\leq \|\mathscr{C} \boxtimes \mathscr{I}(\Gamma) - \Gamma\|_{\mathrm{op}} < \varepsilon,$$

which straightforwardly leads to the thesis. □

Proposition 4.1.3. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and consider*

$$I^{\mathrm{dil}}(\rho) \coloneqq \lim_{\varepsilon\to 0} \limsup_{N\to\infty} \frac{\min\{M : E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N}, \tag{4.14}$$

where $E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$ is the set of the compression schemes $(\mathscr{E}, \mathscr{D})$ such that

$$E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho) = \{(\mathscr{E}, \mathscr{D}) | D^{\mathrm{dil}}(\rho^{\boxtimes N}, \mathscr{C}) < \varepsilon\} \tag{4.15}$$

Then $I(\rho) = I^{\mathrm{dil}}(\rho)$.

Proof. This is a straightforward consequence of lemma 4.1.2. Indeed, all the dilations are a refinement of themselves, so that $E_{N,M,\varepsilon}(\rho) \subseteq E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$ and therefore $I^{\mathrm{dil}}(\rho) \leq I(\rho)$. The inclusion $E_{N,M,\varepsilon}(\rho) \supseteq E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$ follows by lemma 4.1.2, and this implies the thesis. □

In quantum information theory, we have seen that the entanglement fidelity (see equation (2.30)) is the figure of merit that assesses the reliability of a channel in trasmitting entanglement. In a theory that satisfies conditioning and steering, thanks to the above proposition, one can define a quantity analogous to entanglement fidelity by generalizing equation (2.32).

Definition 4.1.4. *Let $\rho \in \mathsf{St}(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A} \to \mathrm{C})$. We define the* correlation fidelity *as follows*

$$F(\rho, \mathscr{C}) = \inf_{\Psi \in D_\rho} F[\Psi, \mathscr{C} \boxtimes \mathscr{I}(\Psi)]^2, \tag{4.16}$$

where the fidelity $F[\Psi, \mathscr{C} \boxtimes \mathscr{I}(\Psi)]$ on the right hand side is that of definition 3.3.1.

The generalised Fuchs-van der Graaf inequalities readily imply the following proposition, showing the equivalence between the fidelity defined in (4.16) and the figure of merit of equation (4.12).

Proposition 4.1.4. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and define*

$$I^F(\rho) \coloneqq \lim_{\varepsilon\to 0} \limsup_{N\to\infty} \frac{\min\{M : E^F_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N},$$

where

$$E^F_{N,M,\varepsilon}(\rho) \coloneqq \{(\mathscr{E}, \mathscr{D}) | F(\rho^{\boxtimes N}, \mathscr{D}\mathscr{E}) > 1 - \varepsilon\},$$

then $I^{\mathrm{dil}}(\rho) = I^F(\rho)$.

Proof. This is a straightforward consequence of proposition 3.3.1. Let $(\mathscr{E}, \mathscr{D}) \in E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$. By the first inequality in (3.46) we have that

$$F[\Psi, (\mathscr{C} \boxtimes \mathscr{I})(\Psi)]^2 \geq 1 - \varepsilon + \frac{\varepsilon^2}{4},$$

for any $\Psi \in D_{\rho^{\boxtimes N}}$, and this implies

$$F(\rho^{\boxtimes N}, \mathscr{C}) > 1 - \varepsilon.$$

Therefore $(\mathscr{E}, \mathscr{D}) \in E^F_{N,M,\varepsilon}(\rho)$, namely $E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho) \subseteq E^F_{N,M,\varepsilon}(\rho)$, whence $I^F(\rho) \leq I^{\mathrm{dil}}(\rho)$.

Now let $(\mathscr{E}, \mathscr{D}) \in E^F_{N,M,\varepsilon}$, then by definition

$$F(\rho^{\boxtimes N}, \mathscr{C}) > 1 - \varepsilon.$$

By the second inequality in proposition 3.3.1 we have that

$$\begin{aligned} &\|(\mathscr{C} \boxtimes \mathscr{I})(\Psi) - \Psi\|_{\mathrm{op}} \\ &\leq 2\sqrt{1 - F[\Psi, (\mathscr{C} \boxtimes \mathscr{I})(\Psi)]^2}, \end{aligned}$$

for any $\Psi \in D_{\rho^{\boxtimes N}}$, which means that $(\mathscr{E}, \mathscr{D}) \in E^{\mathrm{dil}}_{N,M,2\sqrt{\varepsilon}}(\rho)$, and the reverse inequality $I^F(\rho) \geq I^{\mathrm{dil}}(\rho)$ follows. □

4.1.4 The information content in classical and quantum information theory

The reduction of the information content of definition 4.1.3 to Shannon and von Neumann entropy mostly depends on the reduction of our figure of merit to the corresponding cases of classical and quantum information theory. In this respect, conditioning and steering seem to be two key features that cause the boiling down of our definition to the desired cases.

Before considering the quantum case, we prove the following lemma concerning the correlation fidelity defined in 5.1.2 in theories with purification.

Lemma 4.1.3. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{A})$. If every state has a purification (definition 6), then one has*

$$F(\rho, \mathscr{C}) = \inf_{\Phi \in P_\rho} F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2. \tag{4.17}$$

Moreover, in an OPT with essential uniqueness of purification (axiom 7) and purity of parallel composition of states (def. 1), for any $\Phi \in P_\rho$ one has

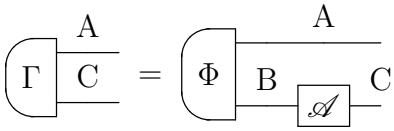
$$F(\rho, \mathscr{C}) = F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2. \tag{4.18}$$

Proof. If $P_\tau \neq \emptyset$ for every state τ, then for any $\Psi \in D_\rho$ one has that there exists $\Gamma \in P_\Psi \subseteq P_\rho$. Therefore, by monotonicity of the fidelity (prop. 3.3.2) we have

$$\begin{aligned} F[\Psi, \mathscr{C} \boxtimes \mathscr{I}(\Psi)]^2 &\geq F[\Gamma, \mathscr{C} \boxtimes \mathscr{I}(\Gamma)]^2 \geq \\ &\geq \inf_{\Phi \in P_\rho} F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2. \end{aligned}$$

Since this holds for any $\Psi \in D_\rho$, it implies $F(\rho, \mathscr{C}) \geq \inf_{\Phi \in P_\rho} F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2$. The reverse inequality is trivial, since $P_\rho \subseteq D_\rho$.

Now, recall that if all the purifications of ρ with the same purifying system are connected through a reversible transformation $\mathscr{U}$ and atomicity of parallel composition of pure states also holds (see property 1 and axiom 7), then, by proposition 3.2.1 for any fixed purification Φ in P_ρ, and any other Γ in P_ρ, there exists a channel $\mathscr{A} \in \mathsf{Transf}_1(\mathrm{B}, \mathrm{C})$ such that



(4.19)

By monotonicity one has $F(\rho, \mathscr{C}) \geq F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2$ and the reverse inequality is trivial, as $\Phi \in P_\rho$. □

The following proposition establishes that, in quantum information theory, the information content is computed via the von Neumann entropy. As we have already said, most of the work has been already done with the simplification of the figure of merit. Indeed, as we have recalled in chapter 3, QT satisfies existence and uniqueness of purification as well as the steering and conditioning properties. Therefore, in QT, the fidelity $F(\rho, \mathscr{C})$ of definition 5.1.2 can be computed according to equation (4.18), where the r.h.s. trivially corresponds to the Uhlmann fidelity. In the proof of the following proposition we see explicitly that the limiting procedure corresponds to the noiseless coding theorem of quantum information theory that we have seen in section 2.2.

Proposition 4.1.5. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ be a quantum state and denote with $S(\rho)$ its von Neumann entropy. Then $I(\rho) = S(\rho)$.*

Proof. We start by showing that $I^F(\rho) \leq S(\rho)$. Let $\delta > 0$ and $\varepsilon > 0$. By the direct part of the Schumacher theorem 2.2.4 there exists a N_0 such that for any $N \geq N_0$ there is a (N, ε)-reliable compression scheme with rate $R \in (S(\rho), S(\rho) + \delta]$. Thus, we have $E^F_{N, NR, \varepsilon}(\rho) \neq \emptyset$ for any $N \geq N_0$. This implies

$$\begin{aligned} &\limsup_{N \to \infty} \frac{\min\{M : E^F_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N} \\ \leq &\limsup_{N \to \infty} \frac{NR}{N} = S(\rho) + \delta. \end{aligned}$$

Since the argument holds for any $\varepsilon > 0$, we get $I^F(\rho) \leq S(\rho) + \delta$, and being delta arbitrary, we find $I^F(\rho) \leq S(\rho)$.

Now let $\delta > 0$ and consider $M/\overline{N}$ such that $S(\rho) - \delta \leq M/\overline{N} < S(\rho)$. By the converse part of Schumacher theorem, given a fixed $\overline{\varepsilon} > 0$, any ε, N-compression scheme with rate $M/\overline{N}$ one has $F(\rho^{\otimes N}, \mathscr{D}\mathscr{E}) \leq 1 - \overline{\varepsilon}$. In particular, since any compression scheme from $k\overline{N}$ copies of the system to kM qubits has ratio $M/\overline{N}$, one has

$$E^F_{k\overline{N}, kM, \overline{\varepsilon}}(\rho) = \emptyset, \quad \forall k \in \mathbb{N}.$$

Therefore, for any $0 < \varepsilon < \overline{\varepsilon}$ and k

$$S(\rho) - \delta \leq \frac{M}{\overline{N}} < \frac{\min\{L : E^F_{k\overline{N},L,\overline{\varepsilon}}(\rho) \neq \emptyset\}}{k\overline{N}} \leq \\ \leq \frac{\min\{L : E^F_{k\overline{N},L,\varepsilon}(\rho) \neq \emptyset\}}{k\overline{N}}.$$

Thus, by taking the $\limsup_{k\to\infty}$ we find that

$$S(\rho) - \delta \leq \limsup_{k\to\infty} \frac{\min\{L : E^F_{k\overline{N},L,\varepsilon}(\rho) \neq \emptyset\}}{k\overline{N}} \\ \leq \limsup_{N\to\infty} \frac{\min\{L : E^F_{N,L,\varepsilon}(\rho) \neq \emptyset\}}{N},$$

for any $0 < \varepsilon < \overline{\varepsilon}$. By taking the $\lim_{\varepsilon\to 0}$ and the arbitrariness of δ we finally get $S(\rho) \leq I^F(\rho)$. The statement then follows by the fact that in quantum theory one has $I^F(\rho) = I(\rho)$ (prop. 4.1.3 and 4.1.4). □

Let us now turn our focus to the classical case. Recall from chapter 2 that, in this setting, the input and the output of the compression scheme are given by strings of N letters drawn from an alphabet X. Each letter x_i appears with a given probability p_i and the probability that the overall string $x_{i_1} \dots x_{i_N}$ is emitted is given by the joint probability $p_{i_1,\dots,i_N}$. If we assume that each symbol is independently and identically distributed, then $p_{i_1,\dots,i_N} = p_{i_1} \dots p_{i_N}$. Now consider *the unique* pure decomposition of some dilation $\mathbf{\Pi} \in D_{\mathbf{p}^{\otimes N}}$. This is given by $\mathbf{\Pi} = \sum_{\mathbf{i},j} \Pi_{ij} \mathbf{e_i} \otimes \mathbf{e}_j$ with $\sum_j \Pi_{\mathbf{i}j} = p_{\mathbf{i}}$ (the pure states of the composite system are the tensor product vectors of the pure ones of the composing systems). Then we find

$$\begin{aligned} &\sum_{\mathbf{i},j} \Pi_{\mathbf{i}j} \|(C \otimes I)\mathbf{e_i} \otimes \mathbf{e}_j - \mathbf{e_i} \otimes \mathbf{e}_j\|_1 \\ &= \sum_{\mathbf{i},j} \Pi_{\mathbf{i}j} \|(C - I)\mathbf{e_i} \otimes \mathbf{e}_j\|_1 \\ &= \sum_{\mathbf{i},j} \Pi_{\mathbf{i}j} \|(C - I)\mathbf{e_i}\|_1 \\ &= \sum_{\mathbf{i}} p_{\mathbf{i}} \|(C - I)\mathbf{e_i}\|_1 = 2p(\mathscr{C}, \rho^{\otimes N}), \end{aligned}$$

having used the fact that for any j

$$\|(C - I)\mathbf{e_i} \otimes \mathbf{e}_j\|_1 = \|(C - I)\mathbf{e_i}\|_1. \tag{4.20}$$

Summarizing, we have proved that for any dilation $\mathbf{\Pi} \in D_{\mathbf{p}^{\otimes N}}$

$$p(\rho, \mathscr{C}) = \frac{1}{2} \sum_{\mathbf{i},j} \Pi_{\mathbf{i}j} \|(C \otimes I)\mathbf{e_i} \otimes \mathbf{e}_j - \mathbf{e_i} \otimes \mathbf{e}_j\|_1. \tag{4.21}$$

Namely, in the classical case, the error probability is equivalent to the figure of merit in equation (4.9). This is enough, since CT satisfies conditioning and every state is proportional to a determinstic one. Notice that the steps outlined above, which have led us to the equality in formula (4.21), are a consequence of two features of CT: the uniqueness of the pure state decomposition, and local discriminability, whereby any bipartite state is separable (according to theorem 3.2.1).

The proof of the fact that the information content, in classical information theory, is computable via the Shannon entropy of the state that represents the source, is completely analogous to the one we have exhibited in the proof of proposition 4.1.5.

4.2 Properties

In the first section we have presented the definition of digitisable theory, that identifies a wide class of theories in which the definition of information content is well-posed, and we have discussed how certain features can help to simplify its computation. In this section we analyse some properties that are shared by the Shannon and von Neumann entropy, such as invariance under reversible operations acting on the system and subaddivity. We also investigate to what extent the information content can be interpreted as a measure of state purity.

4.2.1 First consequences of the definition

Subadditivity is a property of Shannon and von Neumann entropy (item 2 in theorem 2.1.1) and, in turn, a property of information content in the respective theories. It tells us that the information content of a bipartite state is generally smaller than the sum of the information content of marginals. Informally speaking, forgetting about correlations obviously increases uncertainty. For the classical and quantum information content, this property is inherited by the fact that they coincide with the respective entropic functions. Generally, this property does not hold for none the three entropies already introduced in the literature on GPTs (their properties have been discussed in detail in section 3.3) as we will observe in chapter 6, but it holds for the information content as defined in (4.8), independently of how systems are composed in parallel, as shown in the following proposition.

Proposition 4.2.1 (Subadditivity)**.** *Let* $\Psi \in \mathsf{St}_1(\mathrm{AB})$ *and let* $\rho \in \mathsf{St}_1(\mathrm{A})$, $\sigma \in \mathsf{St}_1(\mathrm{B})$ *be its marginals. Then the following property holds*

$$I(\Psi) \leq I(\rho) + I(\sigma).$$

Proof. Let $(\mathscr{E}^\rho, \mathscr{D}^\rho) \in E_{N,\overline{K},\varepsilon}(\rho)$, $(\mathscr{E}^\sigma, \mathscr{D}^\sigma) \in E_{N,\overline{L},\varepsilon}(\sigma)$ with $\overline{K} \coloneqq \min\{M : E_{N,M,\varepsilon}(\rho) \neq \emptyset\}$ and similarly for $\overline{L}$. Now let $\{\Gamma_i\}_{i\in\mathsf{X}}$ be such that $\sum_{i\in\mathsf{X}} \Gamma_i \in D_{\Psi^{\boxtimes N}}$ and consider $\{(\mathscr{C}^\rho \boxtimes \mathscr{I})(\Gamma_i)\}_{i\in\mathsf{X}}$, where $\mathscr{C}^\rho \coloneqq \mathscr{D}^\rho\mathscr{E}^\rho$. Since $\mathscr{C}^\rho$ is a

channel and $D_{\Psi^{\boxtimes N}} \subseteq D_{\rho^{\boxtimes N}}$, $D_{\Psi^{\boxtimes N}} \subseteq D_{\sigma^{\boxtimes N}}$ we have that $\sum_{i\in\mathsf{X}}(\mathscr{C}^\rho \boxtimes \mathscr{I})(\Gamma_i)$ is a dilation of both $\rho^{\boxtimes N}$ and $\sigma^{\boxtimes N}$. Similarly for $\mathscr{C}^\sigma := \mathscr{D}^\sigma\mathscr{E}^\sigma$. This implies the following bound

$$\begin{aligned}
&\sum_{i\in\mathsf{X}}\|(\mathscr{C}^\rho \boxtimes \mathscr{C}^\sigma \boxtimes \mathscr{I})(\Gamma_i) - \Gamma_i\|_{op} \\
=&\sum_{i\in\mathsf{X}}\|(\mathscr{C}^\rho \boxtimes \mathscr{C}^\sigma \boxtimes \mathscr{I})(\Gamma_i) - (\mathscr{C}^\rho \boxtimes \mathscr{I})(\Gamma_i) \\
&+ (\mathscr{C}^\rho \boxtimes \mathscr{I})(\Gamma_i) - \Gamma_i\|_{op} \\
<& 2\varepsilon,
\end{aligned}$$

where we used the triangle inequality for the operational norm. Thus

$$E_{N,\overline{K}+\overline{L},2\varepsilon}(\Psi) \neq \emptyset, \tag{4.22}$$

and this implies that

$$\frac{\min\{M : E_{N,M,2\varepsilon}(\Psi) \neq \emptyset\}}{N} \leq \frac{\overline{K}}{N} + \frac{\overline{L}}{N}.$$

Finally, by taking the $\limsup_{N\to\infty}$ and then $\lim_{\varepsilon\to 0}$ on both sides we get the thesis. □

In classical and quantum theories it is true that the information content is additive (because of the additivity of the respective entropic functions). It is still an open question whether additivity on factorised states , i.e. states of the form $\rho \boxtimes \sigma$, is a general property.

In quantum information theory, another property of the information content due to the von Neumann entropy is the invariance under reversible operation. This is due to the fact that they leave the spectrum untouched. This turns out to be true in full generality, as it is easily proved in the following porposition.

Proposition 4.2.2. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and $\mathscr{U} \in \mathsf{Transf}_1(\mathrm{A})$ be a reversible channel, then $I(\rho) = I(\mathscr{U}(\rho))$.*

Proof. We show that $E_{N,M,\varepsilon}(\mathscr{U}(\rho)) \neq \emptyset \implies E_{N,M,\varepsilon}(\rho) \neq \emptyset$. Let $(\mathscr{E},\mathscr{D}) \in E_{N,M,\varepsilon}(\mathscr{U}(\rho))$ and let $\{\Psi_i\}_{i\in\mathsf{X}}$ be such that $\sum_{i\in\mathsf{X}}\Psi_i \in D_{\rho^{\boxtimes N}}$. It is clear that $\sum_{i\in\mathsf{X}}(\mathscr{U}^{\boxtimes N} \boxtimes \mathscr{I})(\Psi_i) \in D_{\mathscr{U}(\rho)^{\boxtimes N}}$ and therefore

$$\sum_{i\in\mathsf{X}}\|[(\mathscr{D}\mathscr{E} - \mathscr{I}) \boxtimes \mathscr{I}](\mathscr{U}^{\boxtimes N} \boxtimes \mathscr{I})(\Psi_i)\|_{\mathrm{op}} < \varepsilon.$$

Upon defining $\tilde{\mathscr{E}} := \mathscr{E}\mathscr{U}^{\boxtimes N}$ and $\tilde{\mathscr{D}} := (\mathscr{U}^{-1})^{\boxtimes N}\mathscr{D}$, recalling that $\mathscr{U}$ is reversible and that the operational norm is invariant under reversible transformations, the above inequality can be rewritten as follows

$$\begin{aligned}
\varepsilon >& \sum_{i\in\mathsf{X}}\|(\mathscr{U}^{\boxtimes N} \boxtimes \mathscr{I})[(\tilde{\mathscr{D}}\tilde{\mathscr{E}} \boxtimes \mathscr{I})(\Psi_i) - (\Psi_i)]\|_{\mathrm{op}} \\
=& \sum_{i\in\mathsf{X}}\|[(\tilde{\mathscr{D}}\tilde{\mathscr{E}} \boxtimes \mathscr{I})(\Psi_i) - (\Psi_i)]\|_{\mathrm{op}},
\end{aligned}$$

namely, since $\{\Psi_i\}_{i\in\mathsf{X}}$ is arbitrary, $(\tilde{\mathscr{E}},\tilde{\mathscr{D}})\in E_{N,M,\varepsilon}(\rho)\neq\emptyset$. This implies that $R_{N,\varepsilon}(\rho)\leq R_{N,\varepsilon}(\mathscr{U}(\rho))$, and then $I(\rho)\leq I[\mathscr{U}(\rho)]$. The reverse inequality is now trivial

$$I(\rho)=I[\mathscr{U}^{-1}\mathscr{U}(\rho)]\geq I[\mathscr{U}(\rho)],$$

where we have used the previous result along with the fact that $\mathscr{U}^{-1}$ is also reversible. □

4.2.2 Information content versus pure states

Information content in classical and quantum information theory has the property that it is vanishing if and only if the state is pure. Here we see that this correspondence may not be generally true for all OPTs.

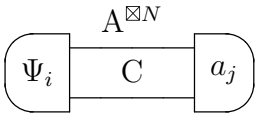
In the following we will assume that $(\mathscr{E},\mathscr{D})$ is a compression scheme with M obits,

$$\mathscr{E}:\mathrm{A}^{\boxtimes N}\to\mathrm{B}^{\boxtimes M},\quad \mathscr{D}:\mathrm{B}^{\boxtimes M}\to\mathrm{A}^{\boxtimes N}.$$

We will denote by $\{\Psi_i\}$ any preparation test of $\mathrm{A}^{\boxtimes N}\mathrm{C}$ such that

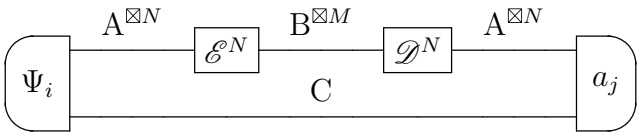
$$\sum_i (e|_{\mathrm{C}}|\Psi_i)_{\mathrm{A}^{\boxtimes N}\mathrm{C}}=\rho^{\boxtimes N}.$$

Finally, for every observation-test $\{a_j\}$ of $\mathrm{A}^{\boxtimes N}\mathrm{C}$, and for any pair $(\mathscr{E},\mathscr{D})$, let us define the two probability distributions

$$p_{i,j}\coloneqq$$ 

(4.23)

and

$$q_{i,j}\coloneqq$$ 

. (4.24)

We can then introduce the following functions that represent the Shannon mutual information between classical random variables X and Y, distributed according to $P(X=x_i,Y=y_j)=p_{i,j}$ or X and $\tilde{Y}$, distributed according to $P(X=x_i,\tilde{Y}=y_j)=q_{i,j}$

$$H(X:Y)\coloneqq\sum_{i,j}p_{i,j}\log_2\frac{p_{i,j}}{p_i^X p_j^Y}$$
$$H(X:\tilde{Y})\coloneqq\sum_{i,j}q_{i,j}\log_2\frac{q_{i,j}}{q_i^X q_j^{\tilde{Y}}},$$

where p_j^Y, $q_j^{\tilde{Y}}$ and $q_i^X=p_i^X$ denote the elements of the marginal distributions.

Definition 4.2.1. *Let $\{\Psi_i\}$ denote a preparation test such that $\sum_i \Psi_i \in D_{\rho^{\boxtimes N}}$. We denote by $E_{N,M,\delta}(\rho)$ the set of those compression schemes such that*

$$\sup_{\mathrm{C},\{\Psi_i\},\{a_j\}} L^{-1}|H(X:Y) - H(X:\tilde{Y})| < \delta,$$

with $L = \log_2(mn-1)$ where n is the cardinality of the preparation test $\{\Psi_i\}$ and m that of the test $\{a_j\}$. We then define the following quantities

$$R^C_{\delta,N}(\rho) \coloneqq \frac{\min\{M \mid E^C_{N,M,\delta}(\rho) \neq \emptyset\}}{N}, \tag{4.25}$$

$$R^C_{\delta}(\rho) \coloneqq \limsup_{N\to\infty} R^C_{\delta,N}(\rho), \tag{4.26}$$

$$I^C(\rho) \coloneqq \lim_{\delta\to 0} R^C_{\delta}(\rho). \tag{4.27}$$

The last quantity above satisfies the two following lemmas.

Lemma 4.2.1. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$. Then $I(\rho) \geq I^C(\rho)$.*

The proof can be found in appendix A. In proving the second lemma and the subsequent proposition we assume that when we compose systems, the size does not increase more than exponentially. More precisely, we formulate the following assumption that will hold in the remainder.

Assumption 2 (Regular scaling)**.** *For every type of system* A*, there exist constants $k_{\mathrm{A}} > 0$ and $D > 0$ such that the size $D(N) \coloneqq D_{\mathrm{A}^{\boxtimes N}}$ of the compound system $\mathrm{A}^{\boxtimes N}$ satisfies $D(N) \leq k_{\mathrm{A}} D^N$ for any N.*

Lemma 4.2.2. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ be a mixed state, then $I^C(\rho) > 0$.*

The proof can be found in appendix B. The proof of this lemma also shows the following bound between $I(\rho)$ and the optimised accessible information defined as in equation (3.49)

$$I(\rho) \geq \frac{S_2(\rho)}{\log_2 D}, \tag{4.28}$$

where D is the constant introduced in the regular scaling assumption 2, where A in this case is the obit system chosen for the definition of $I(\rho)$. This bound is saturated in classical and quantum information theory (where $D = 2$), as well as in fermionic information theory, as we will see in chapter 5. In chapter 6 we will see that in BCT such inequality holds without being saturated.

Proposition 4.2.3. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$. If $I(\rho) = 0$ then ρ is a pure state.*

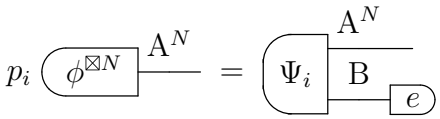
Proof. Let ρ be mixed. By 4.2.1 and 4.2.2

$$I(\rho) \geq I^C(\rho) > 0,$$

whence the thesis. □

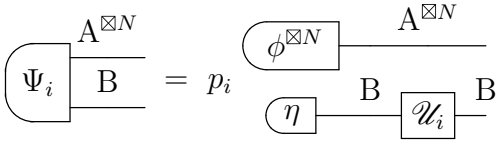
Proposition 4.2.4. *Let us consider strongly causal OPT satisfying regular scaling. Then the requirements of essential uniqueness of purification and purity of parallel composition of states imply that $I(\phi) = 0$ for any $\phi \in \mathsf{PurSt}(\mathrm{A})$. Conversely, if $I(\phi) = 0$ for any $\phi \in \mathsf{PurSt}(\mathrm{A})$, purity of parallel composition of states holds.*

Proof. By proposition 4.1.2 we have $I(\phi) = I^{\mathrm{pur}}(\phi)$. Now, let us fix N and consider a dilation Ω of $\phi^{\boxtimes N}$. Let $\{\Psi_i\}_{i\in\mathsf{X}}$ be a pure decomposition of Ω, then by purity of $\phi^{\boxtimes N}$ we must have



$$\forall i \in \mathsf{X}. \tag{4.29}$$

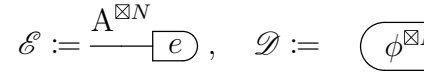
Now, let $\eta \in \mathsf{PurSt}(\mathrm{B})$ and consider $\phi^{\boxtimes N} \boxtimes \eta$. This is still a pure state, hence a purification of $\phi^{\boxtimes N}$. Therefore, by essential uniqueness of purifications



$$\forall i \in \mathsf{X}, \tag{4.30}$$

where $\mathscr{U}_i$ are reversible channels on B.

Now let us consider a compression scheme defined by a measure and prepare one, as follows



$$\tag{4.31}$$

It is clear that for any dilation Ω of $\phi^{\boxtimes N}$, the above scheme is such that $(\mathscr{D}\mathscr{E} \boxtimes \mathscr{I})(\Psi_i) - \Psi_i = 0$ and this implies that for any N and ε we have $E^{\mathrm{pur}}_{N,0,\varepsilon}(\rho) \neq \emptyset$ and then $I(\rho) = I^{\mathrm{pur}}(\rho) = 0$.

Now, let us assume that for any A and for any $\phi \in \mathsf{PurSt}(\mathrm{A})$ we have $I(\phi) = 0$. Let $\rho \in \mathsf{PurSt}(\mathrm{A})$ and $\sigma \in \mathsf{PurSt}(\mathrm{B})$. By proposition 4.2.1 we have that

$$I(\rho \boxtimes \sigma) \leq I(\rho) + I(\sigma) = 0.$$

Thus $I(\rho \boxtimes \sigma) = 0$, and by proposition 4.2.3 $\rho \boxtimes \sigma$ is pure, namely, item 2 of property 1 holds. □

The above result establishes a non trivial relation between purity of states and their informational-theoretic interpretation. On the one hand, in almost full generality, whenever a state has vanishing information content it must be pure, or, in other words, any mixed state has strictly positive information content. On the other hand, sufficient conditions to have null information content on all the pure states are essential uniqueness of purification and purity of parallel composition of states. Remarkably, in any theory where all pure states $\phi \in \mathsf{PurSt}(\mathrm{A})$ for any system A are such that $I(\phi) = 0$, parallel composition must be purity-preserving. When we will compute the information content in BCT we will see explicitly that in a theory violating item 2 of property 1 pure states do not necessarily have vanishing information content.

CHAPTER 5

The information content in Fermionic Theory

From the previous chapter, we understand that a crucial point in a compression protocol is to quantify the reliability of the codec map $\mathscr{C} := \mathscr{D}\mathscr{E}$, which in the asymptotic limit of $N \to \infty$ must coincide with the identity map. In QT checking the reliability of the codec map looking only at its local action, namely via the input-output average fidelity between states $\mathscr{C}(|\phi_{\mathbf{i}}\rangle\langle\phi_{\mathbf{i}}|)$ and $|\phi_{\mathbf{i}}\rangle\langle\phi_{\mathbf{i}}|$, or at the effects on correlations, namely via entanglement fidelity, is equivalent. In addition, we know that QT satisfies the local process tomography, namely that given a map $\mathscr{C}$ on system A one has

$$\begin{aligned} (\mathscr{C}\otimes\mathscr{I}_{\mathrm{C}})(\Psi) =& \Psi \qquad \forall \Psi \in \mathsf{St}(\mathrm{AC}) \\ \Leftrightarrow & \\ \mathscr{C}(\rho) =& \rho \qquad \forall \rho \in \mathsf{St}(\mathrm{A}). \end{aligned} \tag{5.1}$$

This equivalence is due to local discriminability of QT, where the discrimination of bipartite quantum states can always be performed using local measurements only (see in particular equation (3.23)). However, in the absence of local discriminability, a map preserving local states still can affect correlations with remote systems [52]. This raises a crucial issue if one aims at studying the compression task beyond quantum theory, where the reliability of a protocol generally needs to be verified on extended systems. Indeed, in general, testing a compression scheme using ancillary systems is strictly stronger than testing them with local schemes.

As a first example of an information theory without local discriminability, in this chapter we consider the case of fermionic systems as carriers of information. Fermionic computation has been proposed in Ref. [84] and later studied in several works [85, 86, 87, 73, 88]. Differently from quantum systems, fermions obey the *parity superselection rule*. As a consequence, fermionic information theory does not satisfy local discriminability, thus providing a physically relevant example of a theory where the task of compression is not straightforward. Indeed, in the case of study, a map $\mathscr{C}$ that acts as the identity on local states $\rho^{\otimes N}$ could still destroy the correlations with remote systems, and then be mistakenly considered as a reliable codec map.

Exploiting the properties of fermionic quantum information, we prove that the entanglement fidelity is a valid criterion for the reliability of a fermionic codec map. We then show an analogous of the quantum source coding theorem in the fermionic scenario, showing that the minimal compression rate for which a reliable compression scheme exists is the von Neumann entropy of the fermionic state. We conclude therefore that the von Neumann entropy provides the informational content of the state also in the case of fermionic theory, namely in the presence of parity superselection.

5.1 Preliminary considerations

The goal of this chapter is to analyse the behaviour of the information content $I(\rho)$ of a fermionic source of information, and the claim is that it equals the von Neumann entropy of a fermionic state. To prove this claim, we rely on the Jordan-Wigner isomorphism defined in (3.35), in order to take advantage of the Schumacher scheme, which turns out to be a legitimate fermionic protocol, as we will see later. Therefore, in this section, we first define all the information-theoretic quantities in the fermionic setting via their Jordan-Wigner representatives, to exploit the quantum machinery reviewed in section 2.2, and we must first check their independence of the Jordan-Wigner isomorphism.

5.1.1 Setting the problem and the fermionic von Neumann entropy

The main difference between fermionic and quantum information lies in the notion of what Kraus operators correspond to local maps. While in the case of qubit systems local maps acting on the i-th qubit of a composite system have Kraus operators that can be factorized as a non trivial operator on the i-th tensor factor $\mathbb{C}^2$ of the total Hilbert space, in the case of the fermionic Fock space $\mathcal{F}_L$ a local transformation on the i-th mode can be represented in the Jordan-Wigner isomorphism by operators that act non trivially on factors $\mathbb{C}^2$ different from the i-th one. This fact is the source of all the differences between the theory of qubits and fermionic theory, including superselection

and features that it affects, such as the notion of entanglement [60] and local states discrimination protocols [89, 90]. Due to parity superselection, fermionic theory does not satisfy local process tomography, namely the property stating that two transformations $\mathscr{C}_1, \mathscr{C}_2 \in \mathsf{Transf}(\mathrm{L_F} \to \mathrm{M_F})$ are equal iff they act in the same way on the local states in $\mathsf{St}(\mathrm{L_F})$, namely $\mathscr{C}_1(\rho) = \mathscr{C}_2(\rho)$ for every $\rho \in \mathsf{St}(\mathrm{L_F})$. As a consequence, fermionic theory also violates local tomography (see proposition 3.1.1).

Remark 5.1.1. A typical example of a transformation that is locally equivalent to the identity but differs from it when extended to multipartite systems is the parity transformation, as shown in the following. Let us consider a single fermionic mode system 1_F, whose possible states are constrained to be of the form $J(\rho) = q_0 \left|0\right\rangle\left\langle 0\right| + q_1 \left|1\right\rangle\left\langle 1\right|$ by the parity superselection rule. Let P_0 and P_1 be the projectors on $\left|0\right\rangle$ and $\left|1\right\rangle$ respectively, namely on the even and odd sector of the Fock space. The parity transformation $\mathscr{P}$, that in the Jordan-Wigner representation $J(\mathscr{P})$ has Kraus operators P_0 and P_1, acts as the identity $\mathscr{I}_{1_\mathrm{F}}$ when applied to states in $\mathsf{St}(1_\mathrm{F})$. However, taking the system 2_F and considering the extended transformation $\mathscr{P} \boxtimes \mathscr{I}_{1_\mathrm{F}}$ on $\mathsf{St}(2_\mathrm{F})$ one notices that $\mathscr{P}$ differs from the identity map $\mathscr{I}_{1_\mathrm{F}}$. Indeed, the state $J^{-1}(\left|\Psi\right\rangle\left\langle\Psi\right|)$, with $\left|\Psi\right\rangle = \frac{1}{\sqrt{2}}(\left|00\right\rangle + \left|11\right\rangle)$ is a legitimate fermionic state in $\mathsf{St}(2_\mathrm{F})$, and one can straightforwardly verify that

$$\begin{aligned}(\mathscr{P} \boxtimes \mathscr{I}_{1_\mathrm{F}})[J^{-1}(\left|\Psi\right\rangle\left\langle\Psi\right|)] &= \frac{1}{2} J^{-1}(\left|00\right\rangle\left\langle 00\right| + \left|11\right\rangle\left\langle 11\right|) \\ &\neq J^{-1}(\left|\Psi\right\rangle\left\langle\Psi\right|).\end{aligned}$$

That being said, in order to tackle the fermionic noiseless coding problem we consider a top-down approach, that hinges upon the setting described in chapter 4, which also embraces the case of theories that do not satisfy local discriminability.

Let $\mathrm{L_F}$ be a fermionic system and let $\rho \in \mathsf{St}_1(\mathrm{L_F})$ be a generic state of such system. According to the definitions of the previous chapter, the source of fermionic information is supposed to emit N independent copies of the state ρ. Notice that FT is trivially digitisable, since there is essentially a unique system, corresponding to 1_F, and all the others can be thought of as a composition of it. A fermionic compression scheme $(\mathscr{E}_N, \mathscr{D}_N)$ consists of the following two steps:

1. Encoding: Alice encodes the system $\mathrm{L_F}^{\boxtimes N}$ via a channel $\mathscr{E}_N : \mathsf{St}(\mathrm{L_F}^{\boxtimes N}) \to \mathsf{St}(\mathrm{M_F})$, where the target system is generally a system of M-LFMs. The map $\mathscr{E}_N$ produces a fermionic state $\mathscr{E}(\rho^{\boxtimes N})$ with support $\mathsf{Supp}(\mathscr{E}(\rho^{\boxtimes N}))$ on a Fock space $\mathcal{F}_M$ of dimension $d_{\mathcal{F}_M}(N)$ smaller than the one of the original state $\rho^{\boxtimes N}$. The compression rate is defined in the obvious way as the number of modes divided by the length of the message

$$R := \frac{M}{N} = \frac{\log_2 d_{\mathcal{F}_M}(N)}{N}.$$

Alice sends the system $\mathrm{M_F}$ to Bob using the M noiseless fermionic channels.

2. Decoding: Finally Bob sends the system $\mathrm{M_F}$ through a decompression channel $\mathscr{D}_N : \mathsf{St}(\mathrm{M_F}) \to \mathsf{St}(\mathrm{L_F^{\boxtimes N}})$.

The scheme $(\mathscr{E}_N, \mathscr{D}_N)$ overall transforms the $L^{\boxtimes N}$ LFMs, with a codec map $\mathscr{C}_N := \mathscr{D}_N \mathscr{E}_N$. The latter can be more or less "good", in the sense specified by definition 4.1.2, at preserving the information which is contained in the system, depending on ρ itself.

We start by defining the von Neumann entropy of a fermionic state $\rho \in \mathsf{St}_1(\mathrm{L_F})$ in terms of the Jordan-Wigner representative $J(\rho)$ and we then prove the well-posedeness of this definition, namely the independence of the particular isomorphism J.

Definition 5.1.1. *Given a fermionic state ρ, its von-Neumann entropy is defined as*

$$S_f(\rho) := S(J(\rho)) = -\operatorname{Tr}(J(\rho)\log_2 J(\rho)). \tag{5.2}$$

Lemma 5.1.1. *For any femionic system $\mathrm{L_F}$ and any $\rho \in \mathsf{St}_1(\mathrm{L_F})$, $S_f(\rho)$ is well defined.*

Proof. Recall that the operator $J(\rho)\log_2 J(\rho)$ for a density matrix $J(\rho)$ is defined in terms of the eignevalues $\{\lambda_i\}$ of $J(\rho)$ as

$$J(\rho)\log_2 J(\rho) := \sum_i \lambda_i \log_2 \lambda_i \, |x_i\rangle \langle x_i| \tag{5.3}$$

with the usual convention that $0\log_2 0 \equiv 0$. It is clear that, if $\tilde{J}$ is the Jordan-Wigner corresponding to a different ordering, then

$$\tilde{J}(\rho)\log_2 \tilde{J}(\rho) = U J(\rho)\log_2 J(\rho) U^\dagger. \tag{5.4}$$

Since U is unitary, it trivially follows that

$$S_f(\rho) := S[J(\rho)] = S[\tilde{J}(\rho)]. \tag{5.5}$$

□

Since it will be useful later, when we will define the fermionic entanglement fidelity, we also introduce the square root of a fermionic state and we prove its consistency.

Lemma 5.1.2. *The square root of a fermionic state $\rho \in \mathsf{St}_1(\rho)$, defined as*

$$\rho^{\frac{1}{2}} := J^{-1}[J(\rho)^{\frac{1}{2}}], \tag{5.6}$$

is always well defined.

Proof. If $\tilde{J}$ is the Jordan-Wigner isomorphism associated to a different ordering, then consider $X := \tilde{J}^{-1}[\tilde{J}(\rho)^{\frac{1}{2}}]$. We can now prove that $X = \rho^{\frac{1}{2}}$ and then independence of the square root from the ordering. Indeed, one has

$$\tilde{J}(X)^2 = \tilde{J}(\rho) = UJ(\rho)U^\dagger, \tag{5.7}$$

with U unitary. It follows that

$$J(\rho) = U^\dagger \tilde{J}(X) U U^\dagger \tilde{J}(X) U = J(X)^2 \implies J(X) = J(\rho)^{\frac{1}{2}}. \tag{5.8}$$

Since J is an isomorphism, by taking J^{-1} we finally get

$$X = J^{-1}[J(\rho^{\frac{1}{2}})] = \rho^{\frac{1}{2}}. \tag{5.9}$$

Notice that $J(\rho^{\frac{1}{2}}) = J(\rho)^{\frac{1}{2}}$. □

5.1.2 Entanglement fidelity

In chapter 4 we have set a formalism within we have furnished a proper operational definition of $I(\rho)$, in terms of a figure of merit that suitably generalises the error probability of classical information theory and the entanglement fidelity of quantum information theory. Precisely, what we want to do is to compute $I(\rho)$ as defined in equation (4.8) in terms of the fermionic von Neumann entropy $S_f(\rho)$. This task turns out to be hugely simplified, thanks to the fact that FT satisfies both the steering property[1] and conditioning, therefore, $I(\rho) = I^{\mathrm{F}}(\rho)$ according to proposition 4.1.4. On top of that, FT also satisfies existence and uniqueness of purification (axioms 6 and 7) as we now prove. As a consequence of this fact, the fidelity $F(\rho, \mathscr{C})$ is further simplified to equation (4.18), according to lemma 4.1.3. First, we need to prove the following two lemmas.

Lemma 5.1.3. *Consider a quantum state $\rho \in \mathsf{St}(\mathcal{H}_L)$ and two purifications $\Psi, \Phi \in \mathsf{St}(\mathcal{H}_L\mathcal{H}_M)$ with definite parity. Then it is alway possible to find a unitary channel $\mathscr{U}$ that maps states of definite parity into states of definite parity and such that $(\mathscr{I} \otimes \mathscr{U})(\Psi) = \Phi$.*

Proof. Let $|\Psi\rangle \in \mathcal{H}_{LM}^p$ and $|\Phi\rangle \in \mathcal{H}_{LM}^q$, for $p, q \in \{0, 1\}$. Since the two states are purification of the same state $\rho \in \mathsf{St}(\mathcal{H}_L)$ their Schmidt decomposition can always be taken as follows

$$|\Psi\rangle = \sum_i \lambda_i |i\rangle |\Psi_i\rangle, \qquad |\Phi\rangle = \sum_i \lambda_i |i\rangle |\Phi_i\rangle,$$

where $\{|i\rangle\} \in \mathcal{H}_L$ is the same orthonormal set for the two states, while $\{|\Psi_i\rangle\}, \{|\Phi_i\rangle\} \in \mathcal{H}_M$ are two generally different orthonormal sets. Notice

[1] This is a consequence of the fact that FT trivially satisfies the perfect disciminability axiom 2.

that, since Ψ and Φ are pure states of definite parity, any element in the above orthonormal sets must be a vector of definite parity. Within the set $\{|i\rangle\} = \{\{|i_0\rangle\}, \{|i_1\rangle\}\}$ one can separate even $\{|i_0\rangle\}$ and odd $\{|i_1\rangle\}$ parity vectors, and then write Ψ and Φ (respectively of parity p and q) as

$$|\Psi\rangle = \sum_{i_0} \lambda_{i_0} |i_0\rangle \left|\Psi^p_{i_0}\right\rangle + \sum_{i_1} \lambda_{i_1} |i_1\rangle \left|\Psi^{\bar{p}}_{i_1}\right\rangle,$$
$$|\Phi\rangle = \sum_{i_0} \lambda_{i_0} |i_0\rangle \left|\Phi^q_{i_0}\right\rangle + \sum_{i_1} \lambda_{i_1} |i_1\rangle \left|\Phi^{\bar{q}}_{i_1}\right\rangle,$$

where $\bar{r} = r \oplus 1$, and in the orthonormal sets $\{\left|\Psi^p_{i_0}\right\rangle, \left|\Psi^{\bar{p}}_{i_1}\right\rangle\}$ and $\{\left|\Phi^q_{i_0}\right\rangle, \left|\Phi^{\bar{q}}_{i_1}\right\rangle\}$ we separated vectors according to their parity. We can now complete the above two sets to orthonormal bases in such a way that all vectors in both bases have definite parity. Let us take for example the basis $\{\left|\Psi^p_{i_0}\right\rangle, \left|\Psi^{\bar{p}}_{i_1}\right\rangle\}, |\Psi^{r(k)}_k\rangle\}$ and $\{\left|\Phi^1_{i_0}\right\rangle, \left|\Phi^{\bar{q}}_{i_1}\right\rangle, |\Phi^{t(k)}_k\rangle\}$ with $r(k), t(k) \in \{0, 1\}$. It is now straightforward to see that the unitary map $\mathscr{U}$ having Kraus operator

$$U = \sum_{i_0} \left|\Psi^p_{i_0}\right\rangle \left\langle\Phi^q_{i_0}\right| + \sum_{i_1} \left|\Psi^{\bar{p}}_{i_1}\right\rangle \left\langle\Phi^{\bar{q}}_{i_1}\right| + \sum_k |\Psi^{r(k)}_k\rangle\langle\Phi^{t(k)}_k|$$

is such that $(I \otimes U) |\Psi\rangle = |\Phi\rangle$. Moreover $\mathscr{U}$ maps states of definite parity into states of definite parity. □

Lemma 5.1.4. *Let* $\mathrm{N_F} := \mathrm{L_F K_F}$ *and* $\mathscr{C} \in \mathsf{Transf}(\mathrm{N_F} \to \mathrm{N_F})$ *be a single Kraus transformation with Kraus operator* C *having Jordan-Wigner rapresentative* $J(C) = U \otimes I_{\mathrm{K}_F}$, U *acting on the first* L *qubits. Then* $\mathscr{C}$ *is local on the first* L *modes.*

Proof. Due to Proposition 3.2.2, the Kraus operator of $\mathscr{C}$ can be written as $C = \sum_i C_i$, where either each C_i is a product of an even number of field operators, or each C_i is a product of an odd one. The set $\{C_i\}$ can be taken to be linearly independent without loss of generality. Let us assume by contradiction that $\mathscr{C}$ is not local on the first L modes. Therefore, since a set of independent operators generating the algebra of the j-th mode is $\{\varphi_j, \varphi_j^\dagger, \varphi_j^\dagger\varphi_j, \varphi_j^\dagger\varphi_j + \varphi_j\varphi_j^\dagger\}$, there exists at least one product C_i that contains one of the factors φ_j, $\varphi_j^\dagger$, or $\varphi_j\varphi_j^\dagger$, for some mode j of the system $\mathrm{K_F}$. Let $j(i)$ be the mode with largest label in the chosen ordering of the $N = L + K$ modes, such that the corresponding factor in the product C_i is not the identity (i.e. $\varphi_j^\dagger\varphi_j + \varphi_j\varphi_j^\dagger$). Accordingly, one has that the Jordan-Wigner representative of C_i is of the form

$$J(C_i) = K \otimes O_{j(i)} \otimes \left(\bigotimes_{l=j(i)+1}^{N} I_l \right),$$

where K is an operator on the first $1, \ldots, j(i) - 1$ qubits, and $O_{j(i)}$ is one of the factors $\sigma^+_{j(i)}, \sigma^-_{j(i)}, \sigma^+_{j(i)}\sigma^-_{j(i)}$ on the j-th qubit. This contradicts the hypothesis on the form of $J(C)$. □

We are now in the position to prove that, also in FT, any state admits a purification that is unique modulo channels on the purifying system.

Proposition 5.1.1 (Purification of states)**.** *For every $\rho \in \mathsf{St}(\mathrm{L_F})$, there exists a purification $\Psi_\rho \in \mathsf{PurSt}(\mathrm{L_F M_F})$ of ρ for some system $\mathrm{M_F}$. Moreover, the purification is unique up to channels on the purifying system: if $\Psi_\rho \in \mathsf{PurSt}(\mathrm{L_F M_F})$ and $\Phi_\rho \in \mathsf{PurSt}(\mathrm{L_F K_F})$ are two purifications of ρ then there exists a channel $\mathscr{V} \in \mathsf{Transf}_1(\mathrm{M_F} \to \mathrm{K_F})$ such that $(\mathscr{I}_{\mathrm{L_F}} \boxtimes \mathscr{V})(\Psi_\rho) = \Phi_\rho$.*

Proof. It can easily be verified that every purification of $\rho \in \mathsf{St}(\mathrm{L_F})$, having even part ρ_e and odd part ρ_o, can be obtained in terms of the minimal one $J^{-1}(|F\rangle\!\rangle\langle\!\langle F|) \in \mathsf{PurSt}(\mathrm{L_F M_F})$, with $F = J(\rho)^{\frac{1}{2}}$, $M = \lceil \log_2 2r \rceil$ and $r = \max(\mathrm{rank}(\rho_e), \mathrm{rank}(\rho_o))$. Now, let $\Psi_\rho \in \mathsf{PurSt}(\mathrm{L_F M_F})$ and $\Phi_\rho \in \mathsf{PurSt}(\mathrm{L_F K_F})$ be two purifications of ρ. If $M = K$, let us choose the ordering defining the Jordan-Wigner isomorphism of Eq. (3.34) in such a way that the modes in the purifying systems $\mathrm{M_F}$ precede the modes of $\mathrm{L_F}$. Then, using the quantum purification theorem, we know that there exists a reversible map $\mathscr{U}$ with unitary Kraus operator U such that $|F_\rho\rangle\!\rangle = (U \otimes I)|P_\rho\rangle\!\rangle$, where

$$|F_\rho\rangle\!\rangle\langle\!\langle F_\rho| = J(\Phi_\rho), \quad |P_\rho\rangle\!\rangle\langle\!\langle P_\rho| = J(\Psi_\rho).$$

The unitary U can be chosen in such a way that $J^{-1}(\mathscr{U})$ is an admissible fermionic map, namely in such a way that it respects the parity superselection rule (see Lemma 5.1.3). Moreover, due to Lemma 5.1.4, $J^{-1}(U \otimes I)$ cannot contain field operators on the modes in $\mathrm{L_F}$, and is then local on the purifying system $\mathrm{K_F}$. Now, let $K > M$. Then, we can consider a pure state ω on the $K - M$ modes and take the parallel composition $\Psi_\rho \boxtimes \omega$. This is still a purification of ρ, and by the previous argument, there exists a reversible channel $\mathscr{U} \in \mathsf{Transf}_1(\mathrm{K_F} \to \mathrm{K_F})$ such that $\Phi_\rho = (\mathscr{I}_{\mathrm{L_F}} \boxtimes \mathscr{U})(\Psi_\rho \boxtimes \omega) = (\mathscr{I}_{\mathrm{L_F}} \boxtimes \mathscr{V})(\Psi_\rho)$ where $\mathscr{V}$ is the channel defined by the sequential composition $\mathscr{V} = \mathscr{U}(\mathscr{I} \boxtimes \omega)$. If $K < M$, we consider $\Phi_\rho \boxtimes \omega$, where ω is any pure state on $N = M - K$ modes system, and we have $\Phi_\rho \boxtimes \omega = (\mathscr{I}_{\mathrm{L_F}} \boxtimes \mathscr{U})(\Psi_\rho)$. Now we discard the additional modes, and the channel connecting the purifications is the sequential composition of $\mathscr{U}$ and the discarding map: $\mathscr{V} := (\mathscr{I}_{\mathrm{K_F}} \boxtimes \mathrm{Tr}^f_{\mathrm{N_F}})\mathscr{U}$. □

We now define the fermionic entanglement fidelity $F_f(\rho, \mathscr{C})$ as follows

Definition 5.1.2 (Fermionic entanglement fidelity)**.** *Let $\rho \in \mathsf{St}_1(\mathrm{L_F})$, $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{L_F} \to \mathrm{M_F})$ and $\Phi_\rho \in \mathsf{PurSt}(\mathrm{L_F K_F})$ be any purification of ρ. The fermionic Uhlmann's fidelity is defined as*

$$F_f(\rho, \sigma) := \mathrm{Tr}[J(\rho^{1/2}\sigma\rho^{1/2})^{1/2}]. \tag{5.10}$$

Accordingly, the fermionic entanglement fidelity between the states $\rho, \sigma \in \mathsf{St}_1(\mathrm{L_F})$ is defined as

$$F_f(\rho, \mathscr{C}) = F_f[\Phi_\rho, (\mathscr{C} \boxtimes \mathscr{I})(\Phi_\rho)]^2. \tag{5.11}$$

A crucial point is that the above formula makes sense if and only if the fidelity $F(\rho,\sigma)$, as defined in equation (3.45), namely as a minimization over fermionic POVMs of the classical fidelity, can be actually computed via equation (5.10). This is ensured by the following proposition.

Proposition 5.1.2. *Let ρ and σ be two fermionic states. The Uhlmann fidelity $F_f(\rho,\sigma)$ of equation* (5.10) *is well defined and it coincides with $F(\rho,\sigma)$ given in definition 3.3.1.*

Proof. We have already proved that the square root of a fermionic state is well defined in lemma 5.1.2. Moreover, since a reordering of the modes corresponds to a unitarily change of basis, the trace is Jordan-Wigner independent, and so is $F_f(\rho,\sigma)$. What is left to prove is that $F_f(\rho,\sigma) = F(\rho,\sigma)$, and this is done by showing that the POVM that achieves the minimum is made of fermionic effects. Indeed, by remark 2.2.3, we know that the POVM such that $F_f(\rho,\sigma) = F(\rho,\sigma)$ is made of the eigenvectors of the following operator

$$M := J(\rho)^{-\frac{1}{2}}\sqrt{J(\rho)^{\frac{1}{2}}J(\sigma)J(\rho)^{\frac{1}{2}}}J(\rho)^{-\frac{1}{2}}. \tag{5.12}$$

This is a positive operator whose eigenvectors $|m\rangle\langle m|$ have definite parity, since ρ and σ are fermionic states, and this entails that $J^{-1}(|m\rangle\langle m|)$ is a linear combination of even products of fermionic field operators. The thesis then follows by corollary 3.2.1 □

This establishes that $F(\rho,\mathscr{C}) = F_f(\rho,\mathscr{C})$. In the following proposition we finally prove that $F_f(\rho,\mathscr{C})$ can be computed with a formula analogous to the quantum one.

Proposition 5.1.3. *Let $\rho \in \mathsf{St}_1(\mathrm{L_F})$, $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{L_F} \to \mathrm{L_F})$ and $\Phi_\rho \in \mathsf{PurSt}(\mathrm{L_F K_F})$ be any purification of ρ. Entanglement fidelity has the following properties.*

1. *$F(\rho,\mathscr{C})$ is independent of the particular choice for the purification Φ_ρ.*
2. *If the ordering is chosen in such a way that the L modes are all before the purifying ones, the following identity holds:*

$$F(\rho,\mathscr{C}) = \sum_i |\operatorname{Tr}[J(\rho)C_i]|^2 \tag{5.13}$$

for arbitrary Kraus decomposition $J(\mathscr{C}) = \sum_i C_i \cdot C_i^\dagger$ of the Jordan-Wigner representative $J(\mathscr{C})$.

Proof. Let $\Phi_\rho \in \mathsf{PurSt}(\mathrm{L_F K_F})$ be a purification of ρ. If we choose the trivial ordering for the LFMs, the Kraus operators of $J(\mathscr{C}\boxtimes\mathscr{I})$ are of the form $C_i \otimes I$. Moreover, since the minimal purification $|F\rangle\!\rangle\langle\!\langle F|$ (introduced in the proof of proposition 5.1.1) and $J(\Phi_\rho)$ both purify the same quantum state, they are connected through an isometry V. Recalling that for quantum states $|\psi\rangle\langle\psi|$

and σ the quantum Uhlmann fidelity is given by $F(|\psi\rangle\langle\psi|, \sigma) = \langle\psi|\sigma|\psi\rangle^{1/2}$ (equation (2.26)), we find

$$\begin{aligned} F(\rho, \mathscr{C}) &= \sum_i \mathrm{Tr}(|FV^T\rangle\!\rangle\langle\!\langle FV^T|C_i FV^T\rangle\!\rangle\langle\!\langle C_i FV^T|) \\ &= \sum_i |\,\mathrm{Tr}(|C_i FV^T\rangle\!\rangle\langle\!\langle FV^T|)|^2 = \\ &= \sum_i |\,\mathrm{Tr}[J(\rho)C_i]|^2, \end{aligned}$$

namely, the claimed formula in (5.13). Since Φ_ρ is arbitrary, this also implies independence from the choice of the purification. □

5.2 The coding theorem

In this section we define the fermionic typical subspace of a state $\rho \in \mathsf{St}_1(\mathrm{L_F})$ as the typical subspace of the Jordan-Wigner representative $J(\rho)$, and we then prove the noiseless coding theorem for fermionic information theory. The key point is that the typical projector is an admissible kraus operator for a fermionic map, and this fact enables us to use the Schumacher quantum protocol.

5.2.1 Fermionic typical subspaces

When we use the orthonormal decomposition for $J(\rho) = \sum_{x_i} p_i |x_i\rangle\langle x_i|$, this reduces to the Shannon entropy of the classical random variable X that takes values in $\mathsf{Rng}(X) = \{x_1, x_2, \ldots x_n\}$, called range of X, with probability distribution $(p_1, p_2, \ldots, p_n)$: $S_f(\rho) = H(X) = -\sum_i p_i \log_2 p_i$. We remind that N i.i.d. copies of the state ρ are represented as

$$J(\rho^{\boxtimes N}) = J(\rho)^{\otimes N} = \sum_{x_{\mathbf{i}} \in \mathsf{Rng}(X)^N} p_{\mathbf{i}} |x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|. \tag{5.14}$$

With $\mathsf{T}_{N,\varepsilon}(\rho)$ we will denote the typical set of the random variable X.

Definition 5.2.1 (Typical subspace)**.** *Let $\rho \in \mathsf{St}(\mathrm{L_F})$ with orthonormal decomposition $J(\rho) = \sum_{x_i \in \mathsf{Rng}(X)} p_i |x_i\rangle\langle x_i|$. The* δ-typical *subspace $\mathsf{F}_N^\delta(\rho)$ of $\mathcal{H}_L^{\otimes N}$ is defined as*

$$\mathsf{F}_N^\delta(\rho) := \mathsf{Span}\{|x_{\mathbf{i}}\rangle \;\mid\; x_{\mathbf{i}} \in \mathsf{T}_N^\delta(X)\}, \tag{5.15}$$

where $|x_{\mathbf{i}}\rangle := |x_{i_1}\rangle\,|x_{i_2}\rangle \ldots |x_{i_N}\rangle$, and X is the random variable with $\mathsf{Rng}(X) = \{x_i\}$ and $P(x_i) := p_i$.

It is an immediate consequence of the definition of typical subspace that

$$\mathsf{F}_N^\delta(\rho) := \mathsf{Span}\left\{|x_{\mathbf{i}}\rangle \;\mid\; \left|\frac{1}{N}\log_2\frac{1}{p_{\mathbf{i}}} - S_f(\rho)\right| \leq \delta\right\},$$

where $p_{\mathbf{i}} = p_{i_1} \dots p_{i_N}$. We will denote the projector on the typical subspace as

$$\begin{aligned} P_N^\delta(\rho) &:= \sum_{x_{\mathbf{i}} \in \mathsf{T}_N^\delta(X)} |x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}| \\ &= \sum_{x_{\mathbf{i}} \in \mathsf{T}_N^\delta(X)} |x_{i_1}\rangle\langle x_{i_1}| \otimes \cdots \otimes |x_{i_N}\rangle\langle x_{i_N}|, \end{aligned} \tag{5.16}$$

and we have that $\dim(\mathsf{F}_N^\delta(\rho)) = \mathrm{Tr}[P_N^\delta(\rho)] = |\mathsf{T}_N^\delta(X)|$.

Notice that some of the superpositions of vectors in the typical subspace might not be legitimate fermionic pure states, as their parity might be different. However, up to now, we only defined the typical subspace as a mathematical tool, and it does not need a consistent physical interpretation. We will come back to this point later (see Lemma 5.2.1), when we will discuss the physical meaning of the projection $P_{N,\varepsilon}(\rho)$. Now, it is immediate to see that

$$\mathrm{Tr}[P_N^\delta(\rho) J(\rho)^{\otimes N}] = \sum_{x_{\mathbf{i}} \in \mathsf{T}_N^\delta(\mathrm{X})} p_{\mathbf{i}}. \tag{5.17}$$

As in quantum theory, also the fermionic typical subspace has the following features:

Proposition 5.2.1 (Typical subspace)**.** *Let $\rho \in \mathsf{St}(\mathrm{L_F})$. The following statements hold:*

1. *For every $\varepsilon > 0$ and $\delta > 0$ there exists N_0 such that for every $N \geq N_0$*

$$\mathrm{Tr}[P_N^\delta(\rho) J(\rho)^{\otimes N}] \geq 1 - \varepsilon. \tag{5.18}$$

2. *For every $\epsilon > 0$ and $\delta > 0$ there exists N_0 such that for every $N \geq N_0$ the dimension of the typical subspace $\mathsf{F}_N^\delta(\rho)$ is bounded as*

$$(1-\varepsilon) 2^{N(S_f(\rho)-\delta)} \leq \dim(\mathsf{F}_N^\delta(\rho)) \leq 2^{N(S_f(\rho)+\delta)}. \tag{5.19}$$

3. *For given N, let S_N denote an arbitrary orthogonal projection on a subspace of $\mathcal{F}_L^{\otimes N}$ with dimension $\mathrm{Tr}(S_N) < 2^{NR}$, with $R < S_f(\rho)$ fixed. Then for every $\varepsilon > 0$ there exists N_0 such that for every $N \geq N_0$ and every choice of S_N*

$$\mathrm{Tr}[S_N J(\rho)^{\otimes N}] \leq \varepsilon. \tag{5.20}$$

The proof of the above properties is exactly the same as the one of quantum theory (see for instance [26, 27]). However, in order to exploit the same scheme proposed by Schumacher for the quantum case (see also remark 2.2.5), one has to check that the encoding and decoding channels given in the constructive part of the proof are admissible fermionic maps. In particular, the encoding channel makes use of the projector $P_N^\delta(\rho)$ as a Kraus operator, therefore, we have to show that it is a legitimate Kraus operator for a fermionic map. This is proved in the following lemma based on characterization of fermionic transformations of Proposition 3.2.2.

Lemma 5.2.1. *Let ρ be a fermionic state. The projector $P_N^\delta(\rho)$ of equation 5.16 is the Kraus operator of an admissible fermionic transformation.*

Proof. By proposition 3.2.2 the projector on the typical subspace $P_N^\delta(\rho)$ is a legitimate fermionic Kraus operator if it is the sum of products of either an even or an odd number of fermionic fields. Let us consider the single projection $|x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|$. This is given by the tensor product $|x_{i_1}\rangle\langle x_{i_1}| \otimes \cdots \otimes |x_{i_N}\rangle\langle x_{i_N}|$, where each $|x_{i_k}\rangle$ is an eigenvector of the density matrix $J(\rho)$ representing the fermionic state ρ, and, as such, it has a definite parity. Thus, each factor in the above expression of $|x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|$ is the Jordan-Wigner representative of an even polynomial in the field operators, and also the projection $|x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|$ is thus the representative of an even polynomial for every $\mathbf{i}$, which is given, in detail, by the product $J^{-1}(|x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|) = \prod_{j=1}^N J^{-1}(|x_{i_j}\rangle\langle x_{i_j}|)$. Now, by Proposition 3.2.2, $P_N^\delta(\rho)$ is the Jordan-Wigner representative of a legitimate fermionic Kraus operator. □

5.2.2 The fermionic noiseless coding theorem

We can now prove the source coding theorem for fermionic information theory.

Theorem 5.2.1 (Fermionic source coding)**.** *Let $\rho \in \mathsf{St}_1(\mathrm{L_F})$ be a state of system $\mathrm{L_F}$. Then, for any $\delta > 0$ and $\varepsilon > 0$ there exists N_0 such that for any $N \geq N_0$ there exists a comrpession scheme with rate $R \in (S_f(\rho), S_f(\rho) + \delta]$ and $F(\rho^{\boxtimes N}, \mathscr{C}_N) > 1 - \varepsilon$. Moreover, given any rate $R < S_f(\rho)$ and $\varepsilon > 0$ there exists N_0 such that any compression scheme $(\mathscr{E}_N, \mathscr{D}_N)$ with rate R and $N \geq N_0$ is such that $F(\rho^{\boxtimes N}, \mathscr{C}_N) \leq \varepsilon$.*

The proof follows exactly the lines of the original proof for standard quantum compression, that we have discussed in-depth in chapter 2. As the direct proof is constructive, we only need to take care of the legitimacy of the compression protocol as a fermionic map. To this end, we recapitulate the construction here.

1. Encoding: Perform the measurement $\{P_N^{\delta/2}(\rho), I - P_N^{\delta/2}(\rho)\}$. If the outcome corresponding to $P_N^{\delta/2}(\rho)$ occurs, then leave the state unchanged. Otherwise, if the outcome corresponding to $I - P_N^{\delta/2}(\rho)$ occurs, replace the state by a standard state $|S\rangle\langle S|$, with $|S\rangle \in \mathsf{F}_N^{\delta/2}(\rho)$. Such a map is described by the channel $\mathscr{M}_N : \mathrm{L_F}^{\boxtimes N} \to \mathrm{L_F}^{\boxtimes N}$ given by

$$J(\mathscr{M}_N)(\sigma) \coloneqq P_{N,\varepsilon}(\rho)\sigma P_{N,\varepsilon}(\rho) + \mathrm{Tr}[(I - P_{N,\varepsilon}(\rho))\sigma]\,|S\rangle\langle S|\,.$$

   Notice that this is a well defined transformation since by Lemma 5.2.1 the projector on the typical subspace is a legitimate fermionic Kraus operator. The second term is a measure and prepare channel, which is also a legitimate fermionic transformation. Then consider a system $\mathrm{M_F}$ made of M LFMs and

the (partial) isometric embedding $V : \mathsf{F}_N^\delta(\rho) \to \mathcal{H}_M$ such that $V^\dagger V = I_{\mathsf{F}_N^\delta(\rho)}$. Since the first stage of the protocol never produces states in the complement of $\mathsf{F}_N^\delta(\rho)$, we can complete the map $V \cdot V^\dagger$ to a fermionic channel $\mathscr{V}_N$. The encoding is then given by the composite map $\mathscr{E}_N := \mathscr{V}_N \mathscr{M}_N$.

2. Decoding: For the decoding channel, we simply choose the co-isometry $V^\dagger$, which inverts V on $\mathsf{F}_{N,\varepsilon}(\rho)$.

Recall that the fermionic entanglement fidelity $F(\rho^{\boxtimes N}, \mathscr{C}_N)$ can be computed according to formula (5.13). As for the converse statement, the proof for quantum compression is based on item 3, which we proved for fermionic theory as well. Thus, the quantum proof applies to the fermionic case (see also remark 2.2.5).

CHAPTER 6

A case study: Bilocal Classical Theory

As we saw in section 3.3, there are three possible ways of generalising the Shannon and von Neumann entropy as quantifiers of information in OPT, and generally these quantities are inequivalent. A non-trivial issue, at this point, is which of the entropic functions, if any, is an exact quantifier corresponding to the best achievable compression rate of a source. Classical and quantum theory are rather special and fortunate examples of information theories where the three entropic functions coincide, whence they are both monoentropic, and the unique entropy exactly matches the information content of a source, as defined in chapter 4. As we saw in chapter 5, also fermionic information theory is a monoentropic theory where the unique entropy equals the information content. In this scenario, one can then ask two crucial and independent questions: i) what are the features that make a theory monoentropic; ii) under what conditions is the information content quantified by at least one of the three possible entropies—or a regularised version of them.

In this chapter, relying on the minimal framework that we developed in chapter 4, we study the above questions in the Bilocal Classical Theory (BCT) [44] described in subsection 3.2.3. In particular, we show that, while BCT is monoentropic, the information content of its states *does not coincide* with its entropy. Moreover, in BCT pure states actually have non-null information content. The latter feature can be understood considering that the independent preparation of two systems in pure states does not correspond to a pure state for the composite system. Such a preparation thus introduces some ignorance

about the whole system, even if the outcomes of independent experiments on its components are fully predictable. The latter result can be extended to any theory of classical systems where the rule for composing systems is such that the composition of pure states is not necessarily pure. Therefore, thinking of pure states as representing complete knowledge of the physical system is inaccurate if the composition law is not purity-preserving.

6.1 Information content in Bilocal Classical Theory

In this section we prove a noiseless coding theorem in BCT, which in turns yelds a simple formula for the information content of a state in such theory. The result is given in terms of the Shannon entropy of the state describing the source. After having checked that digitisability holds in this theory, thus enabling us to use the apparatus developed in chapter 4, we see how the figure of merit can be simplified thanks to the properties of the theory. Finally, we conclude by showing the main result and discussing some of its consequences.

6.1.1 Checking digitisability

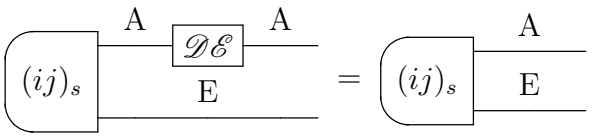
The first fact that must be checked is that BCT is a digitizable theory, in the sense of definition 1. This is a mandatory step that allows for a meaningful operational definition of information content.

Lemma 6.1.1. *BCT is a digitisable theory.*

Proof. We show that any system B of any size D_{B} can serve as obit for BCT. Indeed, let A be any other system of the theory, denote by D_{A} its size and set

$$k = \left\lceil \log_{2D_{\mathrm{B}}} 2D_{\mathrm{A}} \right\rceil . \tag{6.1}$$

Then, let $\mathbf{h} : \{1, \cdots, D_{\mathrm{A}}\} \to \{1, \cdots, D_{\mathrm{B}}\}^k \times \{+, -\}^{k-1}$ be an injiective function. The action of the encoding $\mathscr{E}$ and the decoding $\mathscr{D}$ are defined by two set of probability distributions, $\lambda^{(i)}_{\mathbf{j_s},\tau}$ and $\mu^{(\mathbf{j_s})}_{i,\tau}$ respectively. We then set $\lambda^{(i)}_{\mathbf{h}(i),+} = 1$ $\forall i \in \{1, \cdots, D_{\mathrm{A}}\}$ for the encoding. For the decoding, $\forall \mathbf{j_s}$ such that $\exists! i$ satisfying $\mathbf{h}(i) = \mathbf{j_s}$ we define $\mu^{(\mathbf{j_s})}_{i,+} = 1$, while for every other $\mathbf{j_s}$ we can freely choose any probability distribution. It is now easy to realize that, for any i, j and E, the following holds



Now consider two systems of BCT, say B_1 and B_2. For any integer number k_1 of systems B_1, the minimal number $M_2^{\min}(k_1)$ of B_2 that are needed is given by formula (6.1)

$$M_2^{\min}(k_1) = \left\lceil k_1 \log_{2D_{\mathrm{B}_2}} 2D_{\mathrm{B}_1} \right\rceil ,$$

and similarly

$$M_1^{\min}(k_2) = \left\lceil k_2 \log_{2D_{\mathrm{B}_1}} 2D_{\mathrm{B}_2} \right\rceil .$$

Therefore

$$\lim_{k_1\to\infty} \frac{M_2^{\min}(k_1)}{k_1} = \log_{2D_{\mathrm{B}_2}} 2D_{\mathrm{B}_1} = \left(\lim_{k_2\to\infty} \frac{M_1^{\min}(k_2)}{k_2} \right)^{-1},$$

as required. □

In BCT, as in classical and quantum theory, any type of system can serve as obit. Here we choose the type of system with $D = 2$, that we will call *bibit* from now on. Given that the pure states are the vertices of a simplex, any other state can be uniquely decomposed in terms of pure ones. According to postulate 5, each time we compose in parallel a state with itself, a uniformly distributed sign appears. This entails that a message of length N that is emitted from a source, described by $\sum_i p_i |i)_{\mathrm{A}} = |\rho)_{\mathrm{A}} \in \mathsf{St}_1(\mathrm{A})$, can be written as follows

$$|\rho^{\boxtimes N})_{\mathrm{A}^N} = \sum_{\mathbf{i},\mathbf{s}} p_{\mathbf{i}} \frac{1}{2^{N-1}} |\mathbf{i}_{\mathbf{s}})_{\mathrm{A}^N}, \tag{6.2}$$

where $\mathbf{i}$ and $\mathbf{s}$ collectively denote the string of N local indices and of $N-1$ signs respectively. Notice that, according to the rule of Eq. (3.44), the string of signs $\mathbf{s}$ depends on the order in which the N systems are associated. If the order of composition changes, however, one just has a change in the string of signs $\mathbf{s}' = f(\mathbf{s})$, according to Eq. (3.44), which is immaterial since $\mathbf{s}$ is a dummy index and f is an invertible function. Anyway, for the sake of clarity, we will ubiquitously adopt the convention that the expression in Eq. (6.2) refers to the composite system $(\dots((\mathrm{A}_1\mathrm{A}_2)\mathrm{A}_3)\dots\mathrm{A}_{N-1})\mathrm{A}_N$.

6.1.2 Existence of a mother dilation

The major obstacle in computing the information content of a given state is the complexity of the figure of merit. The greater is the set of states on which we must validate the codec maps, the more difficult is to devise one that works as we wish. However, BCT satisfy both conditioning and steering, and every state is proportional to a deterministic one, so that the figure of merit can be computed according to either equation (4.1.2) or equation (4.12). In other words, the following equations hold

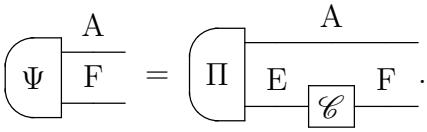
$$I(\rho) = I^{\mathrm{dil}}(\rho) = I^{\mathrm{pur}}(\rho). \tag{6.3}$$

Furthermore, for any state $\rho \in \mathsf{St}_1(\mathrm{A})$ of BCT, there exists a "mother" dilation $\Pi \in \mathsf{St}(\mathrm{AE})$ from which we can compute all the other ones by applying a suitable channel on the ancillary system E, as it is proved in the following proposition.

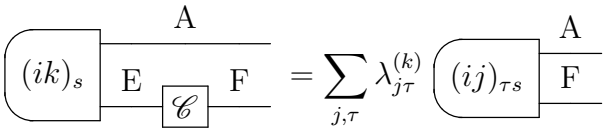
Proposition 6.1.1. *Let* $\sum_i p_i|i)_{\mathrm{A}} = |\rho)_{\mathrm{A}} \in \mathsf{St}_1(\mathrm{A})$ *and let* $\sum_{ijs} q_{ijs}|(ij)_s)_{\mathrm{AF}} = |\Psi)_{\mathrm{AF}} \in \mathsf{St}_1(\mathrm{AF})$ *be a dilation of* ρ*. Let* $\mathrm{E} \cong \mathrm{A}$ *and* $\sum_{iks} p_{iks}|(ik)_s) = |\Pi)_{\mathrm{AE}} \in \mathsf{St}_1(\mathrm{AE})$ *be the dilation of* ρ *with joint probability distribution defined as*

$$p_{iks} := \begin{cases} \delta_{ik} p_i, & s = +; \\ 0, & s = -. \end{cases}$$

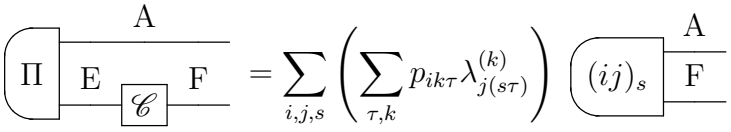
Then, there exists a channel $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{E}, \mathrm{F})$ *such that*



Proof. Any deterministic transformation $\mathscr{C} \in \mathsf{Transf}_1(\mathrm{E}, \mathrm{F})$ acts on the set of pure states of AE as follows



where, for any k, $\lambda^{(k)}_{j\tau}$ is a probability distribution. By applying this to the state $|\Pi)_{\mathrm{AE}}$ defined in the statement, expanding and suitably recollecting the signs, one finds



Since the set of states is a simplex, the pure ones are affinely independent, and this entails that the condition $\mathscr{I}_{\mathrm{A}} \boxtimes \mathscr{C}|\Pi)_{\mathrm{AE}} = |\Psi_{\mathrm{AF}})$ is satisfied iff the following equation holds for any i, j and s

$$q_{ijs} = \sum_{\tau,k} p_{ik\tau} \lambda^{(k)}_{j(s\tau)}. \tag{6.4}$$

For $p_{ik\tau}$ as in the statement

$$q_{ijs} = \sum_{\tau,k} p_{ik\tau} \lambda^{(k)}_{j(s\tau)} = q_{ijs} = \sum_{k} \delta_{ik} p_i \lambda^{(k)}_{js} = p_i \lambda^{(i)}_{js} \tag{6.5}$$

and, since $\sum_{j,s} q_{ijs} = p_i$, $\lambda^{(i)}_{js} = \frac{q_{ijs}}{p_i}$ is an admissible solution. □

The above statement, along with equations (6.3), straightforwardly implies the following proposition, which drastically simplifies the task of devising a compression scheme.

Proposition 6.1.2. *Given a state $\rho \in \mathsf{St}_1(\mathrm{A})$ and a compression scheme $(\mathscr{E}, \mathscr{D})$ for a message of length N, the figure of merit can be computed according to the following formula*

$$\tilde{D}(\rho^{\boxtimes N}, \mathscr{C}) = \sum_{\mathbf{i}} \frac{1}{2^{N-1}} p_{\mathbf{i}} \|(\mathscr{C} - \mathscr{I}_{\mathrm{A}^N}) \boxtimes \mathscr{I}_{\mathrm{A}^N})|(\mathbf{i_s i_s})_+)_{\mathrm{A}^N\mathrm{A}^N}\|_{\mathrm{op}}. \tag{6.6}$$

In other words, defining $\tilde{I}(\rho)$ analogously to $I(\rho)$ (equation (4.8)) by replacing $D(\rho^{\boxtimes N}, \mathscr{C})$ with $\tilde{D}(\rho^{\boxtimes N}, \mathscr{C})$, it holds that $I(\rho) = \tilde{I}(\rho)$.

Proof. Given any state $\rho \in \mathsf{St}_1(\mathrm{A})$, define the following dilation of $\rho^{\boxtimes N}$ (see also equation (6.2)) .

$$|\Pi^N) := \sum_{\mathbf{i},\mathbf{s}} \frac{1}{2^{N-1}} p_{\mathbf{i}} |(\mathbf{i_s i_s})_+)_{\mathrm{A}^N\mathrm{A}^N}. \tag{6.7}$$

Moreover, for the sake of clarity let us set

$$\tilde{E}_{N,M,\varepsilon}(\rho) = \{(\mathscr{E}, \mathscr{D}) | \tilde{D}(\rho^{\boxtimes N}, \mathscr{C}) < \varepsilon\}. \tag{6.8}$$

Denote by $\tilde{I}(\rho)$ the corresponding quantity defined in a way analogous to (4.8). The purpose is to show that $I(\rho) = \tilde{I}(\rho)$. We already know that $I(\rho) = I^{\mathrm{pur}}(\rho) = I^{\mathrm{dil}}(\rho)$. Since it is clear that $\tilde{D}(\rho^{\boxtimes N}, \mathscr{C}) \leq D^{\mathrm{pur}}(\rho^{\boxtimes N}, \mathscr{C})$, it follows that $E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho) \subseteq \tilde{E}_{N,M\varepsilon}(\rho)$, whence $\tilde{I}(\rho) \leq I^{\mathrm{pur}}(\rho) = I(\rho)$. On the other hand, by proposition 6.1.1, lemma 3.1.1 (monotonicity of the operational norm with respect to channels), and the triangle inequality one also has, for any $\Psi \in D_{\rho^{\boxtimes N}}$ and $(\mathscr{E}, \mathscr{D}) \in \tilde{E}_{N,M,\varepsilon}(\rho)$

$$\begin{aligned}
&\|\mathscr{C} \boxtimes \mathscr{I}_{\mathrm{E}}|\Psi)_{\mathrm{A}^N\mathrm{E}} - |\Psi)_{\mathrm{A}^N\mathrm{E}}\|_{\mathrm{op}} = \\
=&\|\mathscr{I}_{\mathrm{A}^N} \boxtimes \mathscr{A}(\mathscr{C} \boxtimes \mathscr{I}_{\mathrm{E}}|\Pi^N)_{\mathrm{A}^N\mathrm{E}} - |\Pi^N)_{\mathrm{A}^N\mathrm{E}})\|_{\mathrm{op}} \\
\leq&\|\mathscr{C} \boxtimes \mathscr{I}_{\mathrm{E}}|\Pi^N)_{\mathrm{A}^N\mathrm{E}} - |\Pi^N)_{\mathrm{A}^N\mathrm{E}}\|_{\mathrm{op}} \leq \\
\leq&\tilde{D}(\rho^{\boxtimes N}, \mathscr{C})
\end{aligned}$$

whence $D^{\mathrm{dil}}(\rho^{\boxtimes N}, \mathscr{C}) \leq \tilde{D}(\rho^{\boxtimes N}, \mathscr{C})$. This implies the inequality $I^{\mathrm{dil}}(\rho) \leq \tilde{I}(\rho)$. We have then proved that $I(\rho) = \tilde{I}(\rho)$. □

At this point, we are in the position to prove a noiseless coding theorem for BCT, whose proof can be found in appendix C.

Theorem 6.1.1. *Let* $\mathrm{A} \in \mathsf{Sys}(\mathrm{BCT})$ *and* $\sum_{i=1}^{D_{\mathrm{A}}} p_i |i)_{\mathrm{A}} = \rho \in \mathsf{St}_1(\mathrm{A})$. *Then*

$$I(\rho) = \frac{H(\mathbf{p}) + 1}{2}. \tag{6.9}$$

A first corollary of formula (6.9) is that the information content is additive in BCT. Indeed, if $\sum_i p_i |i)_{\mathrm{A}} = |\rho)_{\mathrm{A}} \in \mathsf{St}(\mathrm{A})$ and $\sum_j q_j |j)_{\mathrm{B}} = |\sigma)_{\mathrm{B}} \in \mathsf{St}(\mathrm{B})$ then the factorised state is given by

$$|\rho)_{\mathrm{A}} \boxtimes |\sigma)_{\mathrm{B}} = \sum_{i,j,s} \frac{1}{2} p_i q_j |(ij)_s)_{\mathrm{AB}},$$

a straightforward application of equation (6.9) then gives

$$\begin{aligned} I(\rho \boxtimes \sigma) &= \frac{[H(\mathbf{p}) + H(\mathbf{q}) + 1] + 1}{2} \\ &= \frac{H(\mathbf{p}) + 1}{2} + \frac{H(\mathbf{q}) + 1}{2} \\ &= I(\rho) + I(\sigma). \end{aligned}$$

We then deduce that atomicity of parallel composition of states is not a necessary condition for the additivity of information content when independent systems are considered. In particular, not even local discriminability is a necessary condition for additivity of $I(\rho)$. The latter fact, however, was already known from the previous chapter, where we proved that the information content of a fermionic source is given by the von Neumann entropy of the state representing the source.

Another interesting feature of the information content in this theory is that it is strictly positive for all states of any system. In particular, the Shannon entropy of any sharp probability distribution is vanishing, whence, for any system A, it holds that

$$I(\rho) = \frac{1}{2}, \qquad \forall \phi \in \mathsf{PurSt}(\mathrm{A}). \tag{6.10}$$

This is in contrast with what we know from classical, quantum and fermionic theory, where the information content is vanishing if and only if the state is pure. In this respect, one is led to stick to the notion of purity an operational meaning by saying that we have maximal knowledge about a physical system whenever it is in a pure state. In proposition 4.2.4, it has been shown that purity of parallel composition of states and the uniqueness of purifications up to reversible transofrmations on the ancillary system, are sufficient conditions for this interpretation. Notably, a converse is also true, namely that if $I(\phi) = 0$ whenever the state ϕ is pure then state purity is preserved under parallel composition. Here we explicitly see that in a theory with a parallel composition law that is not purity preserving, also pure states can have non vanishing information content.

6.2 Relationship with other entropies

While in the previous section we computed the information content of a state, in the present one we compare it with the three entropies defined in secition 3.3, and with their regularisations. The result finally establishes that, in the OPT framework, none of them can be used as a quantifier of the information content, or, in other words, none of them can hope to satisfy a generalised noiseless coding theorem in the absence of further assumptions on the structure of the theory.

6.2.1 Regularised entropies

Here we consider a regularisation of the three entropic functions obtained by computing such entropies on N copies of a given state, the dividing by N a finally letting N go to infinity. These kind of regularisations are often considered in the literature on quantum information theory [27].

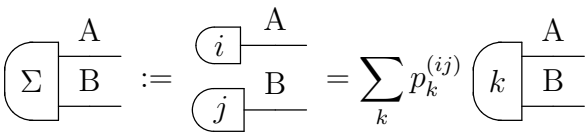
Definition 6.2.1. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$ for some system A. For any $i = 1, 2, 3$ we introduce the regularized S_i^{reg} as follows:*

$$S_i^{\mathrm{reg}}(\rho) := \limsup_{N\to\infty} \frac{S_i(\rho^{\boxtimes N})}{N}. \tag{6.11}$$

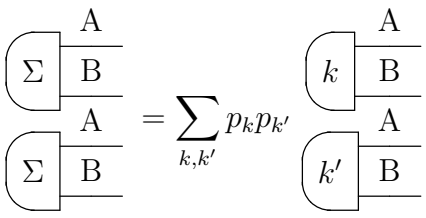
In classical, quantum and fermionic theory, we already know that all the S_i's reduce to the Shannon and von Neumann entropy respectively, and they are all additive, whence they coincide with S_i^{reg} as well. As we have seen in the foregoing section, the relation between the information content of a state and the Shannon entropy of the associated probability distribution is not trivial, as a consequence of the violation of purity of parallel composition of states. The latter property has also a remarkable consequence on the behaviour of the regularized entropies with respect to their single-system counterparts, which is stated in the proposition below.

Proposition 6.2.1. *In any classical theory, for any system A and $\sum_i p_i|i)_{\mathrm{A}} = |\rho)_{\mathrm{A}} \in \mathsf{St}(\mathrm{A})$, one has $S_i(\rho) = H(\mathbf{p})$. Moreover, whenever purity of parallel composition of states (property 1) is violated, there exists a state $\Sigma \in \mathsf{St}(\mathrm{C})$ for some system C such that the strict inequality $S_i^{\mathrm{reg}}(\Sigma) > S_i(\Sigma)$ holds for any $i = 1, 2, 3$.*

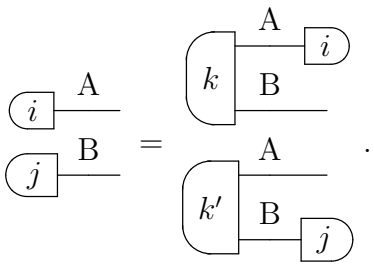
Proof. By hypothesis, there exist systems A and B and $|i)_{\mathrm{A}} \in \mathsf{PurSt}(\mathrm{A})$, $|j)_{\mathrm{B}} \in \mathsf{PurSt}(\mathrm{B})$ such that



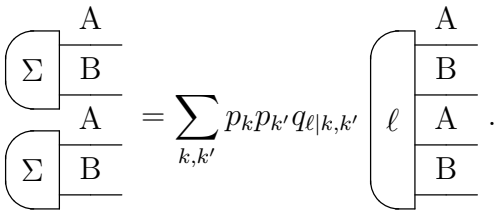
where $\{p_k^{(ij)}\}_{k\in I_{ij}}$ is a non-trivial probability distribution and $\{|k)_{\mathrm{AB}}\}_{k\in I_{ij}}$ is a set of pure states of AB. Now consider $|\Sigma^{\boxtimes 2})_{\mathrm{AB}}$, that can be decomposed as follows



If $|k)_{\mathrm{AB}} \boxtimes |k')_{\mathrm{AB}}$ is pure, then one has a contradiction in that



and, while the left hand side is mixed by hypothesis, by lemma 3.2.1 the right hand side must be pure, since the whole state is pure. We then conclude that $|k)_{\mathrm{AB}} \boxtimes |k')_{\mathrm{AB}}$ must be necessarily mixed. Therefore, $\Sigma^{\boxtimes 2}$ has the following decomposition in terms of pure states



where at least one of the conditioned probability distributions $\{q_{\ell|k,k'}\}_\ell$ is non-trivial. Now, notice that for any system A and any state $|\rho)_{\mathrm{A}} = \sum_i p_i |i)_{\mathrm{A}}$ one has $S_i(\rho) = H(\{p\})$ for any i. The argument that we give here is the same as the one proposed in [54] for classical theory (see in particular proposition 13 and theorem 3(i)), but since states are separating for effects in any OPT it also applies to the case of *any* classical theory defined according to definition 3.2.3. The case of S_3 is pretty obvious, given the uniqueness of the decomposition in terms of pure states. Now, since states are separating for effects, any other atomic effect is proportional to an effect of the perfectly discriminating test, i.e., $(a|_{\mathrm{A}} = \lambda(i|_{\mathrm{A}}$ for some i. Thus, for any state ρ of a classical theory, it follows that any other atomic observation test $\{a_i\}$ is such that $H(a_i(\rho)) \geq H(\mathbf{p})$, by concavity of the function $x \log x$, and the equality is achieved for the perfectly discriminating test, whence $S_1(\rho) = H(\mathbf{p})$. Finally, notice that for any state ρ of a classical theory one has

$$S_2(\rho) = H(\mathbf{p}) - \inf_{\{a_i\} \in \mathcal{O}^{\mathrm{at}}} H(X|J)$$

and $H(X|J) = 0$ for the perfectly discriminating test. Now, a trivial computation shows that

$$\begin{aligned} S_i(\Sigma^{\boxtimes 2}) = & -\sum_{k,k',\ell} p_k p_{k'} q_{\ell|k,k'} \log_2(p_k p_{k'} q_{\ell|k,k'}) = \\ & = 2S_i(\Sigma) - \sum_{k,k'\ell} p_k p_{k'} q_{\ell|k,k'} \log_2(q_{\ell|k,k'}) \\ & > 2S_i(\Sigma). \end{aligned}$$

Finally, notice that the subsequence $S_i(\Sigma^{\boxtimes 2^k})/2^k$ is increasing, since

$$\frac{S_i(\Sigma^{\boxtimes 2^k})}{2^k} \geq \frac{S_i(\Sigma^{\boxtimes 2^{k-1}})}{2^k} + \frac{S_i(\Sigma^{\boxtimes 2^{k-1}})}{2^k} = \frac{S_i(\Sigma^{\boxtimes 2^{k-1}})}{2^{k-1}},$$

and the result follows since the lim sup of the whole sequence is greater than the lim sup of any of its subequence

$$\limsup_{N\to\infty} \frac{S_i(\Sigma^{\boxtimes N})}{N} \geq \lim_{k\to\infty} \frac{S_i(\Sigma^{\boxtimes 2^k})}{2^k} \geq \frac{S_i(\Sigma^{\boxtimes 2})}{2} > S_i(\Sigma).$$

□

As a corollary of the above proposition, we can also notice that, for classical theories where states purity is not preserved, each entropy, in addition to being superadditive, also violates additivity when factorized states are considered. Indeed, if there exist $|i)_{\mathrm{A}} \in \mathsf{PurSt}(\mathrm{A})$ and $|j)_{\mathrm{B}} \in \mathsf{PurSt}(\mathrm{B})$ such that $|i) \boxtimes |j)$ is mixed, we immediately see that

$$S_i(|i) \boxtimes |j)) > 0 = S_i(|i)) + S_i(|j)). \tag{6.12}$$

where S_i is given by the Shannon entropy of the respective decompositions, according to proposition 6.2.1.

6.2.2 Comparison with known entropic functions

While in classical and quantum theory all the S_i's and their regularized version collapse to the Shannon and von Neumann entropies, respectively, thus boiling down to the same operational interpretation given by the noiseless coding theorems, much less is known about their operational meaning in a general theory. In BCT the regularized entropies are related to the Shannon entropy of the state according to the following proposition.

Proposition 6.2.2. *Let* $\sum_i p_i|i)_{\mathrm{A}} = |\rho)_{\mathrm{A}} \in \mathsf{St}_1(\mathrm{A})$*, where* A *is a BCT system. Then* $S_i^{\mathrm{reg}}(\rho) = S_i(\rho) + 1 = H(\mathbf{p}) + 1$.

Proof. BCT is a classical theory, therefore we immediately have that $S_i(\rho) = H(\mathbf{p})$ for any $i = 1, 2, 3$ (by proposition 6.2.1). For S_i^{reg}, we just notice that $S_i(\rho^{\boxtimes N})$ is the Shannon of the factorized joint distribution $p_{\mathbf{i},\mathbf{s}} = p_{\mathbf{i}} \frac{1}{2^{N-1}}$ where $p_{\mathbf{i}} = p_{i_1} \dots p_{i_N}$, thus

$$\frac{S_i(\rho^{\boxtimes N})}{N} = \frac{NH(\mathbf{p}) + (N-1)H(\frac{1}{2})}{N} = H(\mathbf{p}) + 1 - \frac{1}{N}$$

and the result follows by taking the limit. □

A remarkable corollary of this proposition is that, in general, none of the S_i's nor the S_i^{reg}'s can be understood as the minimal compression rate. Notice that, while the additivity property for factorized states is violated by all the entropies, it is satisfied by the regularized versions. The result of proposition 6.2.2 is by far intuitive if we think of the particular compositional rule on states that BCT satisfies. Indeed, at the level of single systems, there is no difference with respect to standard classical theory, and this is true for any classical theory that does not satisfy atomicity of parallel composition of states. The effect of this violation shows up when we consider N copies of the same state, the latter operation giving an extra flat bit, one for each additional copy of the original state, and the appearence of this extra bit is captured by the regularized entropies. The factor 2 in the information content can also be intuitively expected, since also when we compose bibits we get additional space that can be used to allocate the message. Therefore, the departure of $I(\rho)$ from $S_i^{\text{reg}}(\rho)$ can be essentialy ascribed to the weird compositional rule for systems.

We notice that the results of theorem 6.1.1 and proposition 6.2.2 are consistent with the following bound from subsection 4.2.2 (in particular, see the proof of lemma 4.2.2),

$$I(\rho) \geq \frac{S_2(\rho)}{\log_2 D}, \tag{6.13}$$

where D is a costant such that $D_{\mathrm{B}\boxtimes M} \leq kD^M$ for some k. Indeed, in the case of BCT $D_{\mathrm{B}\boxtimes M} = \frac{1}{2}4^M$, whence $D = 4$ (in other words, BCT satisfies the assumption 2 of regular scaling, saturating the inequality). Actually, in the present case such a bound is saturated with $S_2(\rho)$ replaced by $S_2^{\text{reg}}(\rho)$.

Remark 6.2.1. By the result of proposition 6.2.2, in the case of BCT there is a precise relation between the regularized entropy $S_2^{\text{reg}}(\rho)$ and the information content of the following form

$$I(\rho) = \frac{S_2^{\text{reg}}(\rho)}{\log_2 D}, \tag{6.14}$$

where D is the constant such that the relation $D_{\mathrm{B}\boxtimes M} \leq kD^M$ holds for some k. As already noticed, in the case of BCT, this relation is saturated just by taking $k = \frac{1}{2}$ and $D = 4$, whence the equation above trivially holds. One might be tempted to conjecture that a result of this form holds for any classical theory, but it is not difficult to realise that this is not the case. Let us consider a classical theory with only one type of system, say the bit (whose size is 2), satisfying local discriminability and, consequently, purity of parallel composition of states. Now, restrict the allowed tests of the theory to be preparation tests, observation tests, and all possible permutations of bits, when more bits are composed in parallel. It is easy to figure out that there are no protocols that allow one to compress a source represented by a mixed state of a single bit, so that $I(\rho) = 1$. But proposition 6.2.1 implies that $S_i(\rho) = h(p)$ where

$h(p)$ is the binary entropy of the bit state ρ, and so is for $S_i^{\mathrm{reg}}(\rho)$ by local discriminability, whence the conjecture is false.

CHAPTER 7

Conclusions and outlook

The purpose of the present dissertation was to provide a unified framework in which to study the notion of information content, without assuming any feature of classical and quantum theory *ab initio*. To this purpose, the OPT framework enabled us to provide an operational definition of the information content, by lifting the noiseless coding theorems of classical and quantum Shannon theory to a definition . This has been made giving the due care when generalising the compression task in the absence of features such as existence and/or uniqueness of purification, or steering of ensembles. On the one hand, this allows to identify which properties of such a measure of information hold in full generality, and what is the interplay of its features with properties pertaining the fundamental structure of a theory. On the other hand, it lets us to study the behaviour of the information content in toy-theories adopting a top-down approach, that is also helpful to circumscribe a set of conditions that may be sufficient to prove the equivalence between our definition and one of the entropic function introduced in the literature. The present work is a preliminary study of a quantity that can be used to explore the possibility of a reformulation of physical principles in purely informational terms—e.g. a form of holographic principle in context in which space and time are emergent structures—by relying on a notion of entropy that is fully operational.

The first original contribution is in chapter 4, where we defined the information content for a source of information of a nearly arbitrary operational probabilistic theory. The only assumption needed is that of digitisability: a theory is digitisable if any system of the theory can be asymptotically perfectly mapped into finitely many copies of a reference system, called "obit", playing the role

that "bit" and "qubit" play in classical and quantum theory, respectively. The information content of a source is defined as the minimum number of obits needed to store the output of the source in such a way that it can be recovered with arbitrary accuracy. The figure of merit for establishing accuracy, independently of the features of the theory, is robust against any distortion effect that a compression scheme could induce on the state of the source, on its admissible preparations and on the correlations with external systems. Accordingly, the figure of merit meets the following two criteria: i) any preparation of ensembles that average to the considered state must be indistinguishable from leaving the preparation untouched, ii) the compression scheme must preserve decompositions of dilations of the state of interest, namely joint states of the system and arbitrary external systems such that the state that one obtains after averaging and discarding the external system is the one at hand. We first proved that the information content is always a well defined quantity. Moreover, for strongly causal theories satisfying steering of ensembles, we show that the information content can be computed using simple figures of merit, e.g. a generalisation of entanglement fidelity here denoted by correlation fidelity. Then we show that the present notion of information content coincides with the Shannon and von Neumann entropies in the classical and quantum case, respectively. Like Shannon's and von Neumann's entropy, we proved that the information content is subadditive and invariant under reversible transformation, and can be used to measure the purity of a state. Moreover, we investigated to what extent the information content can be understood as a measure of purity on the set of states: while it is always true that a source with null information content corresponds to a pure state, the opposite implication is satisfied in the presence of atomicity of parallel composition (the parallel composition of any two pure states is pure) and unique purification (if a state has a purification, then the latter is unique up to reversible channels on the remote system).

In chapter 5 we studied information compression for fermionic systems, showing the fermionic counterpart of the quantum source coding theorem. In spite of parity superselection rule, which implies strict bilocality, and the nonlocality of the Jordan-Wigner representation of fermionic operators, the von Neumann entropy of fermionic states can still be interpreted as their information content, providing the minimal rate for which a reliable compression is achievable. Given the absence of local discriminability, in FT there exist fermionic admissible maps whose action is equivalent on local states, but they actually differ when local fermionic modes are added and states on extended systems are considered. This suggests that we cannot straightforwardly use the information-theoretic quantities of QT, but it is more appropriate to rely on the apparatus that we established in chapter 4. Remarkably, despite the sensible differences between the quantum and the fermionic case, there are no dissimilarities from the point of view of the source coding.

Finally, in chapter 6 we presented a full computation of the information content in BCT, a strictly bilocal theory where all systems are classical. The result

is given in terms of the Shannon entropy of the probability distribution defining the state, namely, if $|\rho) = \sum_i p_i |i)$, it turns out that $I(\rho) = (H(\mathbf{p}) + 1)/2$. In the special case of BCT, the calculation is simplified by the fact that, for any state, there exists a "mother" dilation from which we can obtain all the other ones by applying a suitable channel on the ancilla. With respect to standard classical Shannon theory, the information content shows two differences that can be both ascribed to the violation of purity of parallel composition of states in BCT: i) there is an overhead given by +1 in the numerator. This is due to the appearance of a bit each time that we compose in parallel a new copy of the same state. Since each bit is uniformly distributed we end up with the maximum overhead, that is indeed +1; ii) there is a factor 2 in the denominator. This follows from the fact that, when we compose bibits into registers, their dimension is given by the formula $D_{\mathrm{B}^{\boxtimes M}} = 2^{M-1} D_{\mathrm{B}}^{M}$, thus the room for allocating messages per single bibit is almost "double" with respect to the size of the register. Notice that the factor 2 in the denominator is then related to meaning of information content in the specific theory at hand, where the elementary systems for physical encoding are bibits. If we had to evaluate the *classical* information content, i.e. the ability of the source to encode classical information, then the regularized mutual information would be the right quantifier. Along with the information content we also analysed the behaviour of three different entropic functions that have been considered in the literature. At the level of single system there is no difference with respect to standard classical theory, and they all coincide with the Shannon entropy of the state ρ. As a consequence, any classical theory is monoentropic [55]. The regularised entropies are clearly sensible to the extra bit that arises when systems are composed, and they all turn out to be equal to $H(\mathbf{p}) + 1$. This result then establishes the existence of theories of information where none of the proposed generalisations of entropy can be interpreted as the information content of the source, and neither their regularized versions do. Remarkably, this is true in a theory that is monoentropic. The departure of the regularised version from the single-system counterpart is not peculiar of the BCT, but it actually takes place in any classical theory (in the sense of the definition 3.2.3) whenever atomicity of parallel composition of states is violated.

Still, there are open questions that are worthy of further investigations. Concerning the general properties of the information content, we noticed that in BCT it is additive on factorized states—i.e. states of the form $\rho \boxtimes \sigma$—and this means that atomicity of parallel composition is not a necessary condition for additivity. A relevant question that is still awaiting an answer is: under which hypotheses, given two states $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$, is it possible to prove that $I(\rho \boxtimes \sigma) = I(\rho) + I(\sigma)$? It might be also the case that additivity is a feature of $I(\rho)$ that holds in full generality, as it would be desirable for measure of the information content. The results of chapter 6 seem to suggest that atomicity of parallel composition plays a marginal role for this property. Moreover, it is left open for further studies what happens if one considers non-local classical theo-

ries with a no-restriction hypothesis on the allowed transformations. It might be the case that a relation very similar in form to equation (6.14) holds. Another interesting question is what happens in the case of the Popescu-Rohrlich boxes [80]. It is possible that one of the entropies is equal to the information content, with a strong prejudice in favor of the optimised mutual information $S_2(\rho)$. On the one hand, we already know that $I(\rho) \geq S_2(\rho)$, therefore the missing part is achievability, i.e. the direct part of a noiseless coding theorem.

The main lesson that we learn from the results presented in this thesis is that a treatment of the notion of information content, from a foundational point of view, cannot ignore the compositional structure of a physical theory, as we have seen that the latter heavily marks its behaviour also in a classical theory. Moreover, as we argued in chapter 6, also the allowed transformations play a significant role, as they might severely restrict the freedom of compressing. We conclude by stressing the relevance of the fundamental questions about information content and entropic functions from the perspective of the extension of area laws or formulations of a general holographic principle, as both these laws rely on entropy, which in turn owes its importance to its operational interpretation as a quantifier of uncertainty (or information content). We think that understanding to what extent these laws can be generalised independently of the nature of information carrying systems can shed new light on the relation between microscopic and large-scale physical phenomena involving the flow of information and its localization [91, 10, 11].

APPENDIX A

Proof of lemma 4.2.1

The following theorem will be useful in the proof of the lemma.

Theorem A.1 ([92], theorem 2)**.** *Let* X,Y *and* X',Y' *be two pairs of random variables taking values on the same range, with joint probability distributions* $p_{i,j}$ *and* $q_{i,j}$ *for* $i = 1, \ldots, m$ *and* $j = 1, \ldots, n$ *respectively, and let* $\gamma := \|\mathbf{p} - \mathbf{q}\|$. *If* $\gamma \leq 1 - \frac{1}{mn}$, *then it holds that*

$$|I(\mathrm{X} : \mathrm{Y}) - I(\mathrm{X}', \mathrm{Y}')| \leq 3\gamma \log_2(mn - 1) + 3H(\gamma). \tag{A.1}$$

Now, we start by defining the following number

$$\zeta(N, \delta) \coloneqq \sup\{\varepsilon \mid E_{N,M,\varepsilon}(\rho) \subseteq E^C_{N,M,\delta}(\rho)\}. \tag{A.2}$$

Firstly, we can observe that in the above definition we can safely take the maximum, since the following inclusion holds

$$E_{N,M,\zeta(N,\delta)}(\rho) \subseteq E^C_{N,M,\delta}(\rho).$$

Indeed, let $(\mathscr{E}, \mathscr{D}) \in E_{N,M,\zeta(N,\delta)}(\rho)$. By definition of $E_{N,M,\zeta(N,\delta)}(\rho)$ we of have

$$\sup_{\mathrm{C},\{\Psi_i\}} \sum_i \|[(\mathscr{D}\mathscr{E} - \mathscr{I}) \boxtimes \mathscr{I}_{\mathrm{C}}]\Psi_i\|_{\mathrm{op}} < \zeta(N, \delta)$$

then there exists $\varepsilon' < \zeta(N, \delta)$ such that $\sup_{\mathrm{C},\{\Psi_i\}} \sum_i \|[(\mathscr{D}\mathscr{E} - \mathscr{I}) \boxtimes \mathscr{I}_{\mathrm{C}}]\Psi_i\|_{\mathrm{op}} < \varepsilon'$. Thus, by definition of $\zeta(N, \delta)$, one has $(\mathscr{E}, \mathscr{D}) \in E_{N,M,\varepsilon''}(\rho)$ with $\varepsilon' < \varepsilon'' < \zeta(N, \delta)$ and $E_{N,M,\varepsilon''}(\rho) \subseteq E^C_{N,M,\delta}(\rho)$. Finally, since $E_{N,M,\varepsilon'}(\rho) \subseteq E_{N,M,\varepsilon''}(\rho)$, we have $(\mathscr{E}, \mathscr{D}) \in E^C_{N,M,\delta}(\rho)$, and consequently $E_{N,M,\zeta(N,\delta)}(\rho) \subseteq E^C_{N,M,\delta}(\rho)$.

This inclusion has another consequence, which is our starting point for proving the lemma. Indeed, by definition one has

$$\limsup_{N\to\infty} R_{\zeta(N,\delta),N}(\rho) \geq R^C_\delta(\rho)$$

We now have the following two possibilities

1. $\exists \delta_0 > 0$ such that, $\forall 0 < \delta < \delta_0$, $\liminf_{N\to\infty} \zeta(N,\delta) = 0$;
2. $\forall \delta > 0$ one has $\liminf_{N\to\infty} \zeta(N,\delta) =: \overline{\zeta}(\delta) > 0$.

Let us start analysing case 2. In this case, by definition of limit inferior, one has

$$\begin{aligned}
&\forall \delta, \gamma > 0 \\
&\begin{cases} \exists N_0, & \forall N \geq N_0, \quad \zeta(N,\delta) > \bar{\zeta}(\delta) - \gamma, \\ \forall N_0, & \exists N \geq N_0, \quad \zeta(N,\delta) < \bar{\zeta}(\delta) + \gamma. \end{cases}
\end{aligned}$$

This implies that for every $\delta > 0$ and every positive γ, for suitably large N it is $R_{\overline{\zeta}(\delta)-\gamma,N}(\rho) \geq R_{\zeta(N,\delta),N}(\rho)$, and consequently, for suitably large N it is $R_{\overline{\zeta}(\delta)/2,N}(\rho) \geq R_{\zeta(N,\delta),N}(\rho)$. In turn, this implies

$$R_{\overline{\zeta}(\delta)/2}(\rho) \geq \limsup_{N\to\infty} R_{\zeta(N,\delta),N}(\rho) \geq R^C_\delta(\rho),$$

and finally, being $\overline{\zeta}(\delta)$ increasing as a function of δ, taking the limit for $\delta \to 0$ one has some value $\varepsilon \geq 0$ such that

$$I(\rho) \geq \lim_{\zeta\to\varepsilon} R_\zeta(\rho) = \lim_{\delta\to 0} R_{\overline{\zeta}(\delta)}(\rho) \geq I^C(\rho).$$

We now turn to case 1, and show that this is not possible. The hypotheses imply indeed that there exists $\delta_0 > 0$ such that $\liminf_{N\to\infty} \zeta(N,\delta_0) = 0$, and the same is then true of every $0 < \delta \leq \delta_0$. This means that for every $\gamma > 0$ and every N_0 there exists $N \geq N_0$ such that $\zeta(N,\delta) < \gamma$ for all $0 < \delta \leq \delta_0$. By definition, this means that for every γ there exists a scheme $(\mathscr{E},\mathscr{D}) \in E_{N,M,\gamma}(\rho)$ such that $(\mathscr{E},\mathscr{D}) \notin E^C_{N,M,\delta}(\rho)$. More explicitly

$$\begin{aligned}
&\sup_{\mathrm{C},\{\psi_i\}} \sum_i \|[(\mathscr{D}\mathscr{E} - \mathscr{I}) \boxtimes \mathscr{I}_{\mathrm{C}}]\psi_i\|_{\mathrm{op}} < \gamma, \\
&\sup_{\mathrm{C},\{\psi_i\},\{a_j\}} L^{-1}|H(\mathrm{X}:\mathrm{Y}) - H(\mathrm{X}:\tilde{\mathrm{Y}})| > \delta,
\end{aligned}$$

where L has been introduced in def. 4.2.1. First of all we remark that if $m = 1$ or $n = 1$, then $H(\mathrm{X}) = 0$ or $H(\mathrm{Y}) = H(\tilde{\mathrm{Y}}) = 0$, respectively, and thus $H(\mathrm{X}:\mathrm{Y}) = H(\mathrm{X}:\tilde{\mathrm{Y}}) = 0$, since $H(\mathrm{A}:\mathrm{B}) \leq \min\{H(\mathrm{A}), H(\mathrm{B})\}$. The

minimum relevant value of L is thus $\log_2 3$. Now according to theorem A.1, for $\|\mathbf{p}-\mathbf{q}\| < \gamma < 1 - 1/mn$ one has

$$\begin{aligned} & L^{-1}|H(\mathrm{X}:\mathrm{Y}) - H(\mathrm{X}':\mathrm{Y}')| \\ \leq & 3\gamma + 3L^{-1}H_2(\gamma) \\ \leq & 3\gamma + \frac{3}{\log_2 3}H_2(\gamma) \end{aligned}$$

where X, Y and X', Y' are distributed according to $p_{i,j}$ and $q_{i,j}$, respectively. We can then conclude that for every $\gamma > 0$ one has

$$\delta < 3\gamma + \frac{3}{\log_2 3}H(\gamma).$$

However, our hypotheses imply that the latter condition must hold for some $\delta > 0$, which is absurd.

APPENDIX B

Proof of Lemma 4.2.2

We premise the following theorem, that will be useful in the proof of the lemma.

Theorem B.1 ([93], theorem 2)**.** *Let* A *be a system of an OPT with size* D_{A}*, and* $\rho \in \mathsf{St}_1(\mathrm{A})$*. Then the maximised accessible information is bounded as follows*

$$S_2(\rho) \leq \log_2(D_{\mathrm{A}}).$$

Let us take $\delta > 0$, and consider $(\mathscr{E}, \mathscr{D}) \in E^C_{N,M,\delta}(\rho)$. Let us consider first a single use of the source associated with ρ corresponding to the decomposition $\{\Psi_i\}$, and let $\{a_j\}$ be the observation test such that $I(\mathrm{X}_0 : \mathrm{Y}_0)$ is maximum, where X_0 is the classical variable corresponding to the outcome i of the preparation test, and Y_0 that of the observation test. Notice that by Krein-Millman's theorem and Caratheodory's theorem one can always find the supremum of mutual information considering atomic decompositions and observation-tests with a bounded number of elements, and thus the optimisation problem has a compact domain. Let now $\{\Psi_{\mathbf{i}}\}$ be the decomposition of $\rho^{\boxtimes N}$ defined by

$$\Psi_{\mathbf{i}} \coloneqq \Psi_{i_1} \boxtimes \Psi_{i_2} \boxtimes \cdots \boxtimes \Psi_{i_N},$$

and $\{\Psi_i\}$ be the decomposition that maximises $I(\mathrm{X}_0 : \mathrm{Y}_0)$, with m_0 outcomes. Let now $\{b_j\}$ be the observation test on N copies of the system that maximizes $H(\mathrm{X} : \mathrm{Y})$ where X is the i.i.d. classical variable given by the preparation event $\mathbf{i}$ and Y by the outcome j. Since $\{b_j\}$ maximizes the mutual information it is clear that the test $\{(b_j|\mathscr{D}\mathscr{E}\}$ will provide a mutual information $H(\mathrm{X} : \tilde{\mathrm{Y}})$ no

larger than $H(\mathrm{X}:\mathrm{Y})$. Thus we can write

$$\begin{aligned}\delta &> \frac{H(\mathrm{X}:\mathrm{Y}) - H(\mathrm{X}:\tilde{\mathrm{Y}})}{\log_2 m_0^N D(N) - 1} \\ &\geq \frac{H(\mathrm{X}:\mathrm{Y}) - H(\mathrm{X}:\tilde{\mathrm{Y}})}{N\log_2 m_0 D_0 + \log_2 k},\end{aligned}$$

where in the first bound we used the fact that the number of outcomes for the observation test maximising the mutual information does not exceed the dimension of the space of effects $D(N)$, while in the second bound we used the hypothesis that there exist k, D_0 such that $D(N) \leq kD_0^N$. Now, by definition of $H(\mathrm{X}:\mathrm{Y})$ we have $H(\mathrm{X}:\mathrm{Y}) \geq NH(\mathrm{X}_0:\mathrm{Y}_0)$, while by the result of theorem B.1 we have

$$H(X:\tilde{Y}) \leq \log_2 D(M) \leq log_2 k' + M\log_2 D_1,$$

where we think of the scheme given by the decomposition $\{\mathscr{E}|\Psi_{\mathbf{i}})\}$ and the observation test given by $\{(b_j|\mathscr{D}\}$, involving M obits. We can then write the following inequality

$$\delta > \frac{NH(X_0:Y_0) - M\log_2 D_1 - \log_2 k'}{N\log_2 m_0 D_0 + \log_2 k},$$

and consequently

$$\begin{aligned}&\frac{M}{N}\frac{\log_2 D_1 + \log_2 k'/M}{\log_2 m_0 D_0 + \log_2 k/N} + \delta \\ &> \frac{H(X_0:Y_0)}{\log_2 m_0 D_0 + \log_2 k/N}.\end{aligned}$$

In particular, if the scheme $(\mathscr{E}, \mathscr{D})$ has the minimum M for fixed N, δ we can then conclude that

$$\begin{aligned}R^C_{\delta,N}\frac{\log_2 D_1}{\log_2 m_0 D_0 + \frac{\log_2 k}{N}} + \delta \\ + \frac{1}{N}\frac{\log_2 k'}{\log_2 m_0 D_0 + \frac{\log_2 k}{N}} \\ > \frac{H(X_0:Y_0)}{\log_2 m_0 D_0 + \frac{\log_2 k}{N}}.\end{aligned}$$

Taking the limit superior for $N \to \infty$ on both sides we have

$$R^C_\delta \frac{\log_2 D_1}{\log_2 m_0 D_0} + \delta \geq \frac{H(X_0:Y_0)}{\log_2 m_0 D_0},$$

and finally, in the limit $\delta \to 0$ we obtain

$$I^C(\rho)\frac{\log_2 D_1}{\log_2 m_0 D_0} \geq \frac{H(X_0:Y_0)}{\log_2 m_0 D_0},$$

namely

$$I^C(\rho) \geq \frac{H(X_0 : Y_0)}{\log_2 D_1}.$$

For a mixed state, $H(X_0 : Y_0) > 0$ and this implies the thesis.

APPENDIX C

Proof of Theorem 6.1.1

We premise the following lemma in order to make the proof of the main theorem a bit less long. The proof makes extensive use of standard techniques from classical information theory.

Lemma C.1. *Let $S(N)$ be any collection of strings $(\mathbf{i}, \mathbf{s})$, with $\mathbf{i} \in \{1, \dots, D_{\mathrm{A}}\}^N$ and $\mathbf{s} \in \{+, -\}^{N-1}$, such that $|S(N)| = 2^{2NR-1}$ with $R < \frac{H(\mathbf{p})+1}{2}$ fixed. Then, for any $\eta > 0$ there exists N_0 such that for any $N \geq N_0$ one has*

$$\sum_{(\mathbf{i},\mathbf{s})\in S(N)} p_{\mathbf{i}} \frac{1}{2^{N-1}} < \eta$$

Proof. Let S be the set of all the possible strings $\mathbf{s}$, $\Delta := H(\mathbf{p}) + 1 - 2R > 0$ and define

$$S_1 := (T^N_{\frac{\Delta}{2}}(\mathbf{p}) \times S) \cap S(N),$$
$$S_2 := (\overline{T^N_{\frac{\Delta}{2}}(\mathbf{p})} \times S) \cap S(N),$$

where $\overline{A}$ denotes the complementary set of A. Then consider

$$\sum_{(\mathbf{i},\mathbf{s})\in S(N)} p_{\mathbf{i}} \frac{1}{2^{N-1}} = \sum_{(\mathbf{i},\mathbf{s})\in S_1} p_{\mathbf{i}} \frac{1}{2^{N-1}} + \sum_{(\mathbf{i},\mathbf{s})\in S_2} p_{\mathbf{i}} \frac{1}{2^{N-1}}.$$

The first term is bounded as follows thanks to the equipartition property in

equation (2.11)

$$\begin{aligned}\sum_{(\mathbf{i},\mathbf{s})\in S_1} p_{\mathbf{i}}\frac{1}{2^{N-1}} &\leq 2^{-N[H(\mathbf{p})-\frac{\Delta}{2}]-N+1}|S(N)| \\ &= 2^{-N[H(\mathbf{p})-\frac{\Delta}{2}]-N+1+2NR-1} \\ &= 2^{-N[H(\mathbf{p})+1-2R-\frac{\Delta}{2}]} < \eta/2\end{aligned}$$

provided that we take $N \geq N_1$ with N_1 sufficiently large. For the term with S_2 we use item 1 of theorem 2.1.2, which implies that for $\eta/2 > 0$ there exists N_2 such that for any $N \geq N_2$ we have $\sum_{\mathbf{i}\in\overline{T^N_{\frac{\Delta}{2}}(\mathbf{p})}} p_{\mathbf{i}} < \frac{\eta}{2}$. We then have

$$\sum_{(\mathbf{i},\mathbf{s})\in S_2} p_{\mathbf{i}}\frac{1}{2^{N-1}} \leq \sum_{\mathbf{i}\in\overline{T^N_{\frac{\Delta}{2}}(\mathbf{p})}} p_{\mathbf{i}} < \frac{\eta}{2}.$$

Setting $N_0 = \max\{N_1, N_2\}$ we have the thesis. □

Proof of theorem 6.1.1. We first prove the achievability, namely that $I(\rho) \leq \frac{H(\{p_i\})+1}{2}$. Let $\delta > 0$, and for any N consider the following number of bibits

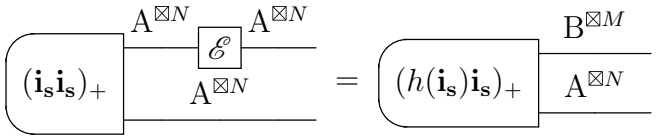
$$M = \left\lceil N\left[\frac{H(\mathbf{p})+1}{2} + \delta\right]\right\rceil$$

by item 2 of theorem 2.1.2 and the above choice of M we have that, $\forall N \in \mathbb{N}$

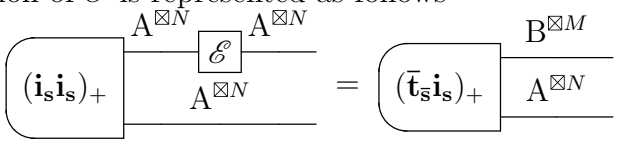
$$|T^N_\delta(\mathbf{p})|2^{N-1} \leq 2^{N[H(\mathbf{p})+1+\delta]-1} \leq 2^{2M-1} = D_{\mathrm{B}^{\boxtimes M}}.$$

This bound entails the existence of a subset $\mathcal{P}_N \subseteq \mathsf{PurSt}(\mathrm{B}^{\boxtimes M})$ with cardinality equal to $|T^N_\delta(\mathbf{p})|2^{N-1}$. Denote by $S_{\mathcal{P}_N}$ the set of strings $\mathbf{t_n}$ associated with $\mathcal{P}_N$. This allows us to define the following $(\mathscr{E}, \mathscr{D})$ compression scheme for any N:

1. The encoding $\mathscr{E}$ is defined by a set of probability distributions $\lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau}$, one for each multi-index $\mathbf{i_s}$. Let $h : T^N_\delta(\mathbf{p}) \times \{+,-\}^{N-1} \to S_{\mathcal{P}_N}$ be any injective function that associates each typical string $\mathbf{i_s}$ with a distinct element $\mathbf{t_n}$ of $S_{\mathcal{P}_N}$. Then, for any such $\mathbf{i_s}$ we set $\lambda^{(\mathbf{i_s})}_{h(\mathbf{i_s})+} = 1$. In particular, for any $\mathbf{i} \in T^N_\delta(\mathbf{p})$ and any $\mathbf{s} \in \{+,-\}^{N-1}$ we have the following diagrammatic equation

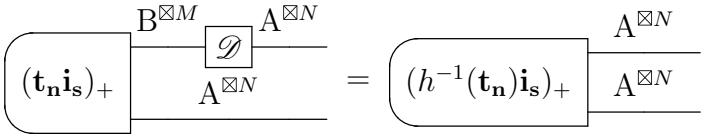


   If $\mathbf{i} \notin T^N_\delta(\mathbf{p})$, for any $\mathbf{s}$ set $\lambda^{(\mathbf{i_s})}_{\bar{\mathbf{t}}_{\bar{\mathbf{s}}}+} = 1$ for a fixed $\bar{\mathbf{t}}_{\bar{\mathbf{s}}} \in S_{\mathcal{P}_N}$. Diagrammatically the action of $\mathscr{E}$ is represented as follows

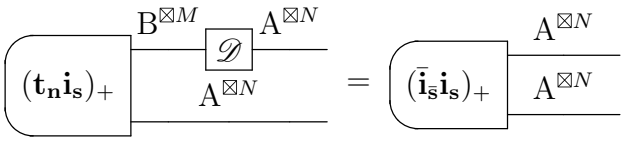


Notice that, having defined $\lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau}$ for any $\mathbf{t_n}$ and τ and for any $\mathbf{i_s}$, we have fully specified the action of $\mathscr{E}$ on all the pure states of $\mathrm{A}^{\boxtimes N}\boxtimes\mathrm{A}^{\boxtimes N}$ even if not all of them directly intervene in the evaluation of the figure of merit.

2. Let $\{\mu^{(\mathbf{t_n})}_{\mathbf{i_s}\tau}\}$ be the probability distributions defining the decoding $\mathscr{D}$. For any $\mathbf{t_n}\in S_{\mathcal{P}_N}$ we simply set $\mu^{(\mathbf{t_n})}_{h^{-1}(\mathbf{t_n})+}=1$, namely, we invert the action of the encoding. Indeed, for any typical string $\mathbf{i_s}=h^{-1}(\mathbf{t_n})$ we have



If $\mathbf{t_n}\notin S_{\mathcal{P}_N}$, take a fixed string $\bar{\mathbf{i}}_{\bar{\mathbf{s}}}$ and define $\mu^{(\mathbf{t_n})}_{\bar{\mathbf{i}}_{\bar{\mathbf{s}}}+}=1$. This implies, for any $\mathbf{i_s}$



Now, with this scheme and using item 1 of theorem 2.1.2, we have that for any $\eta>0$ there exists N_0 such that for any $N\geq N_0$ the following holds

$$
\begin{aligned}
&D(\rho^{\boxtimes N},\mathscr{C})=\\
&\sum_{\mathbf{i}\notin\mathbf{T}^{\mathbf{N}}_{\delta}(\mathbf{p}),\mathbf{s}}\frac{1}{2^{N-1}}p_{\mathbf{i}}\|(|(\bar{\mathbf{i}}_{\bar{\mathbf{s}}}\mathbf{i_s})_+)_{\mathrm{A}^N\mathrm{A}^N}-|\mathbf{i_s}\mathbf{i_s})_+)_{A^NA^N}\|_{\mathrm{op}}\\
&\leq 2\sum_{\mathbf{i}\notin\mathbf{T}^{\mathbf{N}}_{\delta}(\mathbf{p}),\mathbf{s}}\frac{1}{2^{N-1}}p_{\mathbf{i}}\leq 2\sum_{\mathbf{i}\notin\mathbf{T}^{\mathbf{N}}_{\delta}(\mathbf{p})}p_{\mathbf{i}}\leq 2\eta
\end{aligned}
$$

which in turns implies that, given $\epsilon=2\eta$, for any $N\geq N_0$ holds true that $E_{N,M,\epsilon}(\rho)\neq\emptyset$. Therefore, for such values of N we have $M_{N,\epsilon}(\rho)/N\leq M/N$, and this implies, by taking the $\limsup_{N\to\infty}$ and then the $\lim_{\epsilon\to 0}$

$$
I(\rho)\leq\frac{H(\mathbf{p})+1}{2}+\delta,
$$

and the thesis follows by the arbitrariness of δ.

Now we prove the minimality of $\frac{H(\mathbf{p})+1}{2}$, namely that $I(\rho)\geq\frac{H(\mathbf{p})+1}{2}$. Fix an arbitrary $\delta>0$ and let $\overline{M}$ and $\overline{N}$ be such that

$$
\frac{H(\mathbf{p})+1}{2}-\delta\leq\frac{\overline{M}}{\overline{N}}<\frac{H(\mathbf{p})+1}{2}. \tag{C.1}
$$

We then show that there exists N_0 such that for any k satisfying $k\overline{N}>N_0$ it holds that $E_{k\overline{N},k\overline{M},\epsilon}(\rho)=\emptyset$ for any $\epsilon\in(0,\bar{\epsilon}]$ for some $\bar{\epsilon}$. Let $(\mathscr{E},\mathscr{D})$ be a compression scheme for messages of length $k\overline{N}$ with $k\overline{M}$ bibits and let $\lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau}$, $\mu^{(\mathbf{t_n})}_{\mathbf{i_s}\tau}$

be the probability distributions defining the action of $\mathscr{E}$ and $\mathscr{D}$ respectively on pure states of $\mathrm{A}^{\boxtimes N}\mathrm{A}^{\boxtimes N}$. By using the normalization condition the figure of merit can be rewritten in the following form

$$D(\rho^{\boxtimes N}, \mathscr{D}\mathscr{E}) = 2\left(1 - \sum_{\mathbf{i_s}} \frac{1}{2^{N-1}} p_{\mathbf{i}} \sum_{\mathbf{t_n},\tau} \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau} \mu^{(\mathbf{t_n})}_{\mathbf{i_s}\tau}\right). \tag{C.2}$$

Since all the terms $\lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau}\mu^{(\mathbf{t_n})}_{\mathbf{i_s}\tau'}$ are non-negative, we can upper bound the sum over $\tau = \tau'$ by a sum over independent indices τ, τ', obtaining

$$\sum_{\mathbf{i_s}} \frac{1}{2^{N-1}} p_{\mathbf{i}} \sum_{\mathbf{t_n},\tau} \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau} \mu^{(\mathbf{t_n})}_{\mathbf{i_s}\tau} \leq \sum_{\mathbf{i_s}} \frac{1}{2^{N-1}} p_{\mathbf{i}} \sum_{\mathbf{t_n}} \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}} \mu^{(\mathbf{t_n})}_{\mathbf{i_s}}$$

where $\lambda^{(\mathbf{i_s})}_{\mathbf{t_n}} := \sum_\tau \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}\tau}$ and similarly for $\mu^{(\mathbf{t_n})}_{\mathbf{i_s}}$. Now we see that the right hand side of the inequality written above is the sum over a number of strings $\mathbf{i_s}$ at most $2^{2k\overline{M}-1}$; set $p_{\mathbf{i_s}} = p_{\mathbf{i}} \frac{1}{2^{N-1}}$ and consider

$$\sum_{\mathbf{i_s}} p_{\mathbf{i_s}} \sum_{\mathbf{t_n}} \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}} \mu^{(\mathbf{t_n})}_{\mathbf{i_s}}. \tag{C.3}$$

By repeatedly using the inequality $\sum_k a_k b_k \leq \max_k\{b_k\}$ valid for $b_k \geq 0$ and $\sum_k a_k = 1$ we find the following bound

$$\begin{aligned}
\sum_{\mathbf{i_s}} p_{\mathbf{i_s}} \sum_{\mathbf{t_n}} \lambda^{(\mathbf{i_s})}_{\mathbf{t_n}} \mu^{(\mathbf{t_n})}_{\mathbf{i_s}} &\leq \sum_{\mathbf{i_s}} p_{\mathbf{i_s}} \mu^{(\bar{\mathbf{t}}_{\bar{\mathbf{n}}}(\mathbf{i_s}))}_{\mathbf{i_s}} \\
&= \sum_{\mathbf{t_n}} \sum_{\mathbf{i_s}:\bar{\mathbf{t}}_{\bar{\mathbf{n}}}(\mathbf{i_s})=\mathbf{t_n}} p_{\mathbf{i_s}} \mu^{(\bar{\mathbf{t}}_{\bar{\mathbf{n}}}(\mathbf{i_s}))}_{\mathbf{i_s}} \\
&\leq \sum_{\mathbf{t_n}} p_{\mathbf{i_s}(\mathbf{t_n})}
\end{aligned} \tag{C.4}$$

In the last sum we have one $\mathbf{i_s}$ for any $\mathbf{t_n}$, therefore the sum is over at most 2^{2M-1} terms. Notice that our choice of $M_k = k\overline{M}$ entails $2^{2M_k-1} = 2^{2\frac{\overline{M}}{\overline{N}}N_k-1} < 2^{N_k H(\mathbf{p})+N_k-1}$. Thus, by lemma C.1, given $\eta > 0$ there exists N_0 such that for any k satisfying $k\overline{N} > N_0$ we have $\sum_{\mathbf{i_s}} \frac{1}{2^{N-1}} p_{\mathbf{i}} < \eta$. This implies that, for any $\epsilon \in (0, 2(1-\eta)]$

$$E_{k\overline{N},k\overline{M},\epsilon} = \emptyset \qquad \forall k > \left\lceil \frac{N_0}{\overline{N}} \right\rceil$$

Therefore, for any such k, we have the following chain of inequalities

$$\begin{aligned}
\frac{H(\mathbf{p})+1}{2} - \delta &\leq \frac{k\overline{M}}{k\overline{N}} \\
&\leq \limsup_{k\to\infty} \frac{M_{k\overline{N},\epsilon}(\rho)}{k\overline{N}} \\
&\leq \limsup_{N\to\infty} \frac{M_{N,\epsilon}(\rho)}{N}
\end{aligned}$$

Taking the $\lim_{\epsilon\to 0}$ we have $\frac{H(\mathbf{p})+1}{2} - \delta \leq I(\rho)$ and the thesis follows by the arbitrariness of δ. □

Bibliography

[1] Jacob D Bekenstein. Black holes and the second law. *Lettere Al Nuovo Cimento (1971–1985)*, 4(15):737–740, 1972.

[2] Jacob D Bekenstein. Black holes and entropy. *Physical Review D*, 7(8):2333, 1973.

[3] Jacob D. Bekenstein. Generalized second law of thermodynamics in black-hole physics. *Phys. Rev. D*, 9:3292–3300, Jun 1974. URL: https://link.aps.org/doi/10.1103/PhysRevD.9.3292, doi:10.1103/PhysRevD.9.3292.

[4] Jacob D Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Physical Review D*, 23(2):287, 1981.

[5] S. W. Hawking. Gravitational radiation from colliding black holes. *Phys. Rev. Lett.*, 26:1344–1346, May 1971. URL: https://link.aps.org/doi/10.1103/PhysRevLett.26.1344, doi:10.1103/PhysRevLett.26.1344.

[6] S. W. Hawking. Black holes in general relativity. *Commun. in Math. Phys.*, 25:152–166, June 1972. doi:10.1007/BF01877517.

[7] B. Carter. Axisymmetric black hole has only two degrees of freedom. *Phys. Rev. Lett.*, 26:331–333, Feb 1971. URL: https://link.aps.org/doi/10.1103/PhysRevLett.26.331, doi:10.1103/PhysRevLett.26.331.

[8] Leonard Susskind. Lattice fermions. *Phys. Rev. D*, 16:3031–3039, Nov 1977. doi:10.1103/PhysRevD.16.3031.

[9] G. 't Hooft. Dimensional reduction in quantum gravity, 1993. URL: https://arxiv.org/abs/gr-qc/9310026, doi:10.48550/ARXIV.GR-QC/9310026.

[10] Raphael Bousso. The holographic principle. *Rev. Mod. Phys.*, 74:825–874, Aug 2002. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.825, doi:10.1103/RevModPhys.74.825.

[11] Jens Eisert, Marcus Cramer, and Martin B Plenio. Colloquium: Area laws for the entanglement entropy. *Reviews of Modern Physics*, 82(1):277, 2010.

[12] John Archibald Wheeler. *Information, Physics, Quantum: The Search for Links*, pages 354–368. Addison-Wesley, 1990. doi:10.1007/978-94-017-1201-9_4.

[13] G. Birkhoff and J. von Neumann. The logic of quantum mechanics. *Math. Ann.*, 37:823, 1936.

[14] G. Ludwig. *Foundations of quantum mechanics*. Springer-Verlag, 1983.

[15] C. A. Fuchs. Quantum foundations in light of quantum information. In *NATO Science Series, III: Computer and Systems Sciences*, volume 182, pages 38–82, 2001.

[16] L. Hardy. Quantum theory from five reasonable axioms. *Arxiv preprint quant-ph/0101012*, 2001.

[17] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Informational derivation of quantum theory. *Phys. Rev. A*, 84:012311, Jul 2011. URL: https://link.aps.org/doi/10.1103/PhysRevA.84.012311, doi:10.1103/PhysRevA.84.012311.

[18] M. Tribus and E. C. McIrvine. Energy and information. *Scientific American*, 224, 1971.

[19] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[20] Alfréd Rényi. On measures of information and entropy. In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pages 547–561, 1960.

[21] Benjamin Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, Apr 1995. URL: https://link.aps.org/doi/10.1103/PhysRevA.51.2738, doi:10.1103/PhysRevA.51.2738.

[22] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1932.

[23] Richard Jozsa and Benjamin Schumacher. A new proof of the quantum noiseless coding theorem. *Journal of Modern Optics*, 41(12):2343–2349, 1994. arXiv:https://doi.org/10.1080/09500349414552191, doi:10.1080/09500349414552191.

[24] Howard Barnum, Christopher A. Fuchs, Richard Jozsa, and Benjamin Schumacher. General fidelity limit for quantum channels. *Phys. Rev. A*, 54:4707–4711, Dec 1996. URL: https://link.aps.org/doi/10.1103/PhysRevA.54.4707, doi:10.1103/PhysRevA.54.4707.

[25] Te Sun Han and Kingo Kobayashi. *Mathematics of Information and Coding*, volume 203. American Mathematical Society, 2002.

[26] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. doi:10.1017/CBO9780511976667.

[27] Mark M. Wilde. *Quantum Information Theory*. Cambridge University Press, 2013. doi:10.1017/CBO9781139525343.

[28] G.M. D'Ariano, P. Lo Presti, and M.F. Sacchi. Bell measurements and observables. *Physics Letters A*, 272(1):32–38, 2000. URL: https://www.sciencedirect.com/science/article/pii/S0375960100004102, doi:https://doi.org/10.1016/S0375-9601(00)00410-2.

[29] Richard Jozsa. Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12):2315–2323, 1994. arXiv:https://doi.org/10.1080/09500349414552171, doi:10.1080/09500349414552171.

[30] C.A. Fuchs and J. van de Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 1999. doi:10.1109/18.761271.

[31] Benjamin Schumacher. Sending entanglement through noisy quantum channels. *Phys. Rev. A*, 54:2614–2628, Oct 1996. URL: https://link.aps.org/doi/10.1103/PhysRevA.54.2614, doi:10.1103/PhysRevA.54.2614.

[32] Andreas Winter. Coding theorems of quantum information theory, 1999. URL: https://arxiv.org/abs/quant-ph/9907077, doi:10.48550/ARXIV.QUANT-PH/9907077.

[33] Michał Horodecki. Limits for compression of quantum information carried by ensembles of mixed states. *Phys. Rev. A*, 57:3364–3369, May 1998. URL: https://link.aps.org/doi/10.1103/PhysRevA.57.3364, doi:10.1103/PhysRevA.57.3364.

[34] Michal Horodecki. Optimal compression for mixed signal states. *Phys. Rev. A*, 61:052309, Apr 2000. URL: https://link.aps.org/doi/10.1103/PhysRevA.61.052309, doi:10.1103/PhysRevA.61.052309.

[35] Howard Barnum, Carlton M Caves, Christopher A Fuchs, Richard Jozsa, and Benjamin Schumacher. On quantum coding for ensembles of mixed

states. *J. Phys. A: Math. Gen. 34 6767*, 34(35):6767–6785, aug 2001. doi:10.1088/0305-4470/34/35/304.

[36] Masato Koashi and Nobuyuki Imoto. Compressibility of quantum mixed-state signals. *Phys. Rev. Lett.*, 87:017902, Jun 2001. URL: https://link.aps.org/doi/10.1103/PhysRevLett.87.017902, doi:10.1103/PhysRevLett.87.017902.

[37] Zahra Baghali Khanian and Andreas Winter. Entanglement-assisted quantum data compression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1147–1151, 2019. doi:10.1109/ISIT.2019.8849352.

[38] Zahra Baghali Khanian and Andreas Winter. General mixed state quantum data compression with and without entanglement assistance. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1852–1857, 2020. doi:10.1109/ISIT44484.2020.9174148.

[39] Lucien Hardy. Disentangling nonlocality and teleportation, 1999. arXiv:quant-ph/9906123.

[40] Martin Plávala. General probabilistic theories: An introduction, 2021. URL: https://arxiv.org/abs/2103.07469, doi:10.48550/ARXIV.2103.07469.

[41] Jonathan Barrett. Information processing in generalized probabilistic theories. *Phys. Rev. A*, 75:032304, Mar 2007. doi:10.1103/PhysRevA.75.032304.

[42] Robert W. Spekkens. Evidence for the epistemic view of quantum states: A toy theory. *Phys. Rev. A*, 75:032110, Mar 2007. doi:10.1103/PhysRevA.75.032110.

[43] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Probabilistic theories with purification. *Phys. Rev. A*, 81:062348, Jun 2010. doi:10.1103/PhysRevA.81.062348.

[44] Giacomo Mauro D'Ariano, Marco Erba, and Paolo Perinotti. Classicality without local discriminability: Decoupling entanglement and complementarity. *Phys. Rev. A*, 102:052216, Nov 2020. URL: https://link.aps.org/doi/10.1103/PhysRevA.102.052216, doi:10.1103/PhysRevA.102.052216.

[45] Lluís Masanes and Markus P Müller. A derivation of quantum theory from physical requirements. *New Journal of Physics*, 13(6):063001, jun 2011. URL: https://dx.doi.org/10.1088/1367-2630/13/6/063001, doi:10.1088/1367-2630/13/6/063001.

[46] Howard Barnum, Markus P Müller, and Cozmin Ududec. Higher-order interference and single-system postulates characterizing quantum theory. *New Journal of Physics*, 16(12):123029, dec 2014. URL: https://dx.doi.org/10.1088/1367-2630/16/12/123029, doi:10.1088/1367-2630/16/12/123029.

[47] B Dakic and C Brukner. Quantum theory and beyond: is entanglement special? In *Deep Beauty: Understanding the Quantum World through Mathematical Innovation*. Cambridge University Press, 2011.

[48] Philip Goyal, Kevin H. Knuth, and John Skilling. Origin of complex quantum amplitudes and feynman's rules. *Phys. Rev. A*, 81:022109, Feb 2010. URL: https://link.aps.org/doi/10.1103/PhysRevA.81.022109, doi:10.1103/PhysRevA.81.022109.

[49] Alexander Wilce. Conjugates, Filters and Quantum Mechanics. *Quantum*, 3:158, July 2019. doi:10.22331/q-2019-07-08-158.

[50] Howard Barnum, Jonathan Barrett, Matthew Leifer, and Alexander Wilce. Generalized no-broadcasting theorem. *Phys. Rev. Lett.*, 99:240501, Dec 2007. URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.240501, doi:10.1103/PhysRevLett.99.240501.

[51] Howard Barnum, Jonathan Barrett, Matthew Leifer, and Alexander Wilce. Teleportation in General Probabilistic Theories. In *Mathematical Foundations of Information Flow: Proceedings of the Clifford Lectures*, volume 71, pages 25–47, 2008. URL: http://arxiv.org/abs/0805.3553.

[52] Giacomo Mauro D'Ariano, Paolo Perinotti, and Alessandro Tosini. Information and disturbance in operational probabilistic theories. *Quantum*, 4:363, November 2020. doi:10.22331/q-2020-11-16-363.

[53] G M D'Ariano, F Manessi, and P Perinotti. Determinism without causality. *Physica Scripta*, 2014(T163):014013, dec 2014. URL: https://dx.doi.org/10.1088/0031-8949/2014/T163/014013, doi:10.1088/0031-8949/2014/T163/014013.

[54] Gen Kimura, Koji Nuida, and Hideki Imai. Distinguishability measures and entropies for general probabilistic theories. *Reports on Mathematical Physics*, 66(2):175–206, 2010. URL: https://www.sciencedirect.com/science/article/pii/S003448771000025X, doi:https://doi.org/10.1016/S0034-4877(10)00025-X.

[55] Howard Barnum, Jonathan Barrett, Lisa Orloff Clark, Matthew Leifer, Robert Spekkens, Nicholas Stepanik, Alex Wilce, and Robin Wilke. Entropy and information causality in general probabilistic theories. *New Journal of Physics*, 12:3024, mar 2010. doi:10.1088/1367-2630/12/3/033024.

[56] Anthony J Short and Stephanie Wehner. Entropy in general physical theories. *New Journal of Physics*, 12:3023, March 2010. URL: http://dx.doi.org/10.1088/1367-2630/12/3/033023, doi:10.1088/1367-2630/12/3/033023.

[57] Giacomo Mauro D'Ariano, Giulio Chiribella, and Paolo Perinotti. *Quantum theory from first principles: an informational approach*. Cambridge University Press, 2017.

[58] Lucien Hardy and William K Wootters. Limited holism and real-vector-space quantum theory. *Foundations of Physics*, 42(3):454–473, 2012. URL: http://dx.doi.org/10.1007/s10701-011-9616-6, doi:10.1007/s10701-011-9616-6.

[59] Giacomo Mauro D'Ariano, Marco Erba, and Paolo Perinotti. Classical theories with entanglement. *Phys. Rev. A*, 101:042118, Apr 2020. URL: https://link.aps.org/doi/10.1103/PhysRevA.101.042118, doi:10.1103/PhysRevA.101.042118.

[60] Giacomo Mauro D'Ariano, Franco Manessi, Paolo Perinotti, and Alessandro Tosini. The feynman problem and fermionic entanglement: Fermionic theory versus qubit theory. *International Journal of Modern Physics A*, 29(17):1430025, 2014. arXiv:https://doi.org/10.1142/S0217751X14300257, doi:10.1142/S0217751X14300257.

[61] Huzihiro Araki. On a characterization of the state space of quantum mechanics. *Communications in Mathematical Physics*, 75(1):1–24, 1980. doi:10.1007/BF01962588.

[62] Howard Barnum and Alexander Wilce. Local tomography and the jordan structure of quantum theory. *Foundations of Physics*, 44(2):192–212, 2014. doi:10.1007/s10701-014-9777-1.

[63] Paolo Perinotti. Cellular automata in operational probabilistic theories. *Quantum*, 4:294, July 2020. doi:10.22331/q-2020-07-09-294.

[64] E. Schrödinger. Probability relations between separated systems. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(3):446–452, 1936. doi:10.1017/S0305004100019137.

[65] Howard Barnum, Carl Philipp Gaebler, and Alexander Wilce. Ensemble steering, weak self-duality, and the structure of probabilistic theories. *Foundations of Physics*, 43:1411–1427, 2013. arXiv:https://doi.org/10.1142/S0217751X14300257, doi:10.1007/s10701-013-9752-2.

[66] Nicetu Tibau Vidal, Mohit Lal Bera, Arnau Riera, Maciej Lewenstein, and Manabendra Nath Bera. Quantum operations in an information theory for fermions. *Phys. Rev. A*, 104:032411, Sep 2021. URL:

https://link.aps.org/doi/10.1103/PhysRevA.104.032411, doi:10.1103/PhysRevA.104.032411.

[67] Ernst Pascual Jordan and Eugene Paul Wigner. Über das paulische äquivalenzverbot. *Zeitschrift für Physik*, 47(9-10):631–651, September 1928. doi:10.1007/bf01331938.

[68] Frank Verstraete and Juan Ignacio Cirac. Mapping local hamiltonians of fermions to local hamiltonians of spins. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09012–P09012, September 2005. URL: https://doi.org/10.1088%2F1742-5468%2F2005%2F09%2Fp09012, doi:10.1088/1742-5468/2005/09/p09012.

[69] Carlos Pineda, Thomas Barthel, and Jens Eisert. Unitary circuits for strongly correlated fermions. *Physical Review A*, 81(5):050303, May 2010. URL: https://link.aps.org/doi/10.1103/PhysRevA.81.050303, doi:10.1103/physreva.81.050303.

[70] Norbert Schuch, Frank Verstraete, and Juan Ignacio Cirac. Nonlocal resources in the presence of superselection rules. *Physical Review Letters*, 92(8):087904, February 2004. URL: https://link.aps.org/doi/10.1103/PhysRevLett.92.087904, doi:10.1103/physrevlett.92.087904.

[71] Alexei Kitaev, Dominic Mayers, and John Preskill. Superselection rules and quantum protocols. *Physical Review A*, 69(5):052326, May 2004. URL: https://link.aps.org/doi/10.1103/PhysRevA.69.052326, doi:10.1103/physreva.69.052326.

[72] Norbert Schuch, Frank Verstraete, and Juan Ignacio Cirac. Quantum entanglement theory in the presence of superselection rules. *Physical Review A*, 70(4):042310, October 2004. URL: https://link.aps.org/doi/10.1103/PhysRevA.70.042310, doi:10.1103/physreva.70.042310.

[73] Giacomo Mauro D'Ariano, Franco Manessi, Paolo Perinotti, and Alessandro Tosini. Fermionic computation is non-local tomographic and violates monogamy of entanglement. *EPL (Europhysics Letters)*, 107(2):20009, July 2014. doi:10.1209/0295-5075/107/20009.

[74] Giacomo Mauro D'Ariano, Paolo Perinotti, and Alessandro Tosini. Incompatibility of observables, channels and instruments in information theories. *Journal of Physics A: Mathematical and Theoretical*, 55(39):394006, sep 2022. URL: https://dx.doi.org/10.1088/1751-8121/ac88a7, doi:10.1088/1751-8121/ac88a7.

[75] C. Zander and A. R. Plastino. Fidelity measure and conservation of information in general probabilistic theories. *Europhysics Letters*, 86(1):18004, apr 2009. URL: https://dx.doi.org/10.1209/0295-5075/86/18004, doi:10.1209/0295-5075/86/18004.

[76] Gen Kimura, Junji Ishiguro, and Makoto Fukui. Entropies in general probabilistic theories and their application to the holevo bound. *Phys. Rev. A*, 94:042113, Oct 2016. URL: https://link.aps.org/doi/10.1103/PhysRevA.94.042113, doi:10.1103/PhysRevA.94.042113.

[77] Ryo Takakura. Entropy of mixing exists only for classical and quantum-like theories among the regular polygon theories. *Journal of Physics A: Mathematical and Theoretical*, 52(46):465302, oct 2019. URL: https://dx.doi.org/10.1088/1751-8121/ab4a2e, doi:10.1088/1751-8121/ab4a2e.

[78] Ryo Takakura and Takayuki Miyadera. Entropic uncertainty relations in a class of generalized probabilistic theories. *Journal of Physics A: Mathematical and Theoretical*, 54(31):315302, jul 2021. URL: https://dx.doi.org/10.1088/1751-8121/ac0c5c, doi:10.1088/1751-8121/ac0c5c.

[79] Marcin Pawłowski, Tomasz Paterek, Dagomir Kaszlikowski, Valerio Scarani, Andreas Winter, and Marek Żukowski. Information causality as a physical principle. *Nature*, 461(7267):1101–1104, 2009. doi:10.1038/nature08400.

[80] Sandu Popescu and Daniel Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24(3):379–385, Mar 1994. doi:10.1007/BF02058098.

[81] Giacomo Mauro D'Ariano and Alessandro Tosini. Testing axioms for quantum theory on probabilistic toy-theories. *Quantum Information Processing*, 9(2):95–141, 2010.

[82] Gen Kimura, Koji Nuida, and Hideki Imai. Distinguishability measures and entropies for general probabilistic theories. *Reports on Mathematical Physics*, 66(2):175–206, October 2010. URL: http://dx.doi.org/10.1016/S0034-4877(10)00025-X, doi:10.1016/s0034-4877(10)00025-x.

[83] David Gross, Markus Müller, Roger Colbeck, and Oscar C. O. Dahlsten. All reversible dynamics in maximally nonlocal theories are trivial. *Phys. Rev. Lett.*, 104:080402, Feb 2010. URL: https://link.aps.org/doi/10.1103/PhysRevLett.104.080402, doi:10.1103/PhysRevLett.104.080402.

[84] Sergey B. Bravyi and Alexei Yu. Kitaev. Fermionic quantum computation. *Annals of Physics*, 298(1):210 – 226, 2002. doi:10.1006/aphy.2002.6254.

[85] Michael M. Wolf. Violation of the entropic area law for fermions. *Physical Review Letters*, 96(1):010404, January 2006. URL: https://

link.aps.org/doi/10.1103/PhysRevLett.96.010404, doi:10.1103/physrevlett.96.010404.

[86] Mari-Carmen Bañuls, Juan Ignacio Cirac, and Michael M. Wolf. Entanglement in fermionic systems. *Physical Review A*, 76(2):022311, August 2007. URL: https://link.aps.org/doi/10.1103/PhysRevA.76.022311, doi:10.1103/physreva.76.022311.

[87] Nicolai Friis, Antony R. Lee, and David Edward Bruschi. Fermionic-mode entanglement in quantum information. *Physical Review A*, 87(2):022338, February 2013. URL: https://link.aps.org/doi/10.1103/PhysRevA.87.022338, doi:10.1103/physreva.87.022338.

[88] Tiago Debarba, Fernando Iemini, Geza Giedke, and Nicolai Friis. Teleporting quantum information encoded in fermionic modes. *Phys. Rev. A*, 101:052326, May 2020. URL: https://link.aps.org/doi/10.1103/PhysRevA.101.052326, doi:10.1103/PhysRevA.101.052326.

[89] Matteo Lugli, Paolo Perinotti, and Alessandro Tosini. Fermionic state discrimination by local operations and classical communication. *Phys. Rev. Lett.*, 125:110403, Sep 2020. URL: https://link.aps.org/doi/10.1103/PhysRevLett.125.110403, doi:10.1103/PhysRevLett.125.110403.

[90] Matteo Lugli, Paolo Perinotti, and Alessandro Tosini. Unambiguous discrimination of fermionic states through local operations and classical communication. *Phys. Rev. A*, 103:012416, Jan 2021. URL: https://link.aps.org/doi/10.1103/PhysRevA.103.012416, doi:10.1103/PhysRevA.103.012416.

[91] Leonard Susskind. The world as a hologram. *Journal of Mathematical Physics*, 36(11):6377–6396, 1995. arXiv:https://doi.org/10.1063/1.531249, doi:10.1063/1.531249.

[92] Zhengmin Zhang. Estimating mutual information via kolmogorov distance. *IEEE Transactions on Information Theory*, 53(9):3280–3282, 2007. doi:10.1109/TIT.2007.903122.

[93] Samuel Fiorini, Serge Massar, Manas K Patra, and Hans Raj Tiwary. Generalized probabilistic theories and conic extensions of polytopes. *J. Phys. A: Math. Theor.*, 48(2):025302, January 2015. doi:10.1088/1751-8113/48/2/025302.

List of publications

- P. Perinotti, A. Tosini, and L. Vaglini, "Shannon theory beyond quantum: Information content of a source", Phys. Rev. A **105**, 052222 (2022).
- P. Perinotti, A. Tosini, and L. Vaglini, "Shannon theory for quantum systems and beyond: information compression for fermions", arXiv:2106.04964 (2021);
  (accepted for publication in the special issue: The Quantum-Like Revolution: A Festschrift for Andrei Khrennikov, 2023, edited by Springer);
- P. Perinotti, A. Tosini, and L. Vaglini, "Which entropy for general physical theories?", arXiv:2302.01651 (currently under review for the Quantum journal).