Shannon theory beyond quantum: Information content of a source

Paolo Perinotti,* Alessandro Tosini,† and Leonardo Vaglini‡
QUIT Group, Physics Department, Pavia University, and INFN Sezione di Pavia, via Bassi 6, 27100 Pavia, Italy

(Received 31 December 2021; accepted 6 May 2022; published 25 May 2022)

The information content of a source is defined in terms of the minimum number of bits needed to store the output of the source in a perfectly recoverable way. A similar definition can be given in the case of quantum sources, with qubits replacing bits. In the mentioned cases the information content can be quantified through Shannon's and von Neumann's entropy, respectively. Here we extend the definition of information content to operational probabilistic theories, and prove relevant properties such as the subadditivity and the relation between purity and information content of a state. We prove the consistency of the present notion of information content when applied to the classical and the quantum case. Finally, the relation with one of the notions of entropy that can be introduced in general probabilistic theories, the maximum accessible information, is given in terms of a lower bound.

DOI: 10.1103/PhysRevA.105.052222

I. INTRODUCTION

The birth of information theory, marked by Shannon's pioneering work [1], is represented by a thorough definition of information content of an information source, along with its quantification through a suitable quantity—the celebrated Shannon entropy. Since its early times, the information content was identified with the minimum number of elementary information carriers—bits—needed to encode messages from the source in a perfectly recoverable way.

The notion of information content was then lifted to the quantum scenario, where the elementary carrier is a qubit. Schumacher [2] proved that the quantum information content of a quantum source can be quantified through its von Neumann's entropy. The typical setting for the proof of the above mentioned results entails a regularization procedure, i.e., considering an arbitrarily large number of uses of the source, and encoding schemes that are not perfect, but arbitrarily accurate. This definition makes the notion of information content depend on the choice of the figure of merit used to evaluate accuracy [3–5]. While in the classical case one uses the (complement of) error probability, Schumacher's theorem uses the entanglement fidelity, which evaluates how well the encoding scheme preserves entanglement of the source with an external system. As we show here, this figure of merit is equivalent to the average input-output fidelity for an arbitrary decomposition of the state representing the source. One big lesson of quantum information theory is thus that preserving coherence is equivalent to preserving entanglement, in a motto: *quantum information is entanglement* [6].

In a broader sense, the above argument teaches us to judge the action of a transformation looking not only at the way it transforms input states, but also correlations with remote systems. While in quantum theory the two perspectives are somehow interchangeable, there are other information theories where the second one becomes mandatory, such as real quantum theory [7] or fermionic theory [8,9]. The latter two theories are examples of operational probabilistic theories [10–13] (OPTs) or more generally of generalized probabilistic theories [14–16]. The definition of OPTs sets a framework of alternative theories of elementary information carriers (or possibly elementary physical systems) and their processes, that can be thought of as all the conceivable information theories. In the present work we show that it is actually possible to extend the notion of information content discussed above to the framework of OPTs, and compare it to some entropic functions introduced in this context by analogy with known entropies [17–19].

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 was recently discussed in other contexts, e.g., in assessing information vs disturbance [20].

It is worth mentioning one of the main assumptions made in the present work. In classical and quantum Shannon theories the amount of information is measured in bits and qubits respectively, as we have recalled just above. 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 "digitizable." Roughly speaking, we assume the existence of

*paolo.perinotti@unipv.it
†alessandro.tosini@unipv.it
‡leonardo.vaglini01@universitadipavia.it

at least one elementary system, which we call obit, such that an agent can always encode an arbitrary but finite number of copies of her or 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 from a foundational perspective, such as nonlocal boxes [21,22]. 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 [17–19]. Moreover, this assumption can also be applied to theories without local tomography, such as real quantum theory and quantum theory with superselection (e.g., fermionic theory).

After introducing the notion of information content in a general OPT, we prove some of its main properties and show that the optimized accessible information [17–19] generally provides a lower bound for it. As special cases, we then analyze classical and quantum theory, where our definition boils down to Shannon's and von Neumann's entropy, respectively. As a consequence of the present definition, finally, fermionic information content can be proved to coincide with von Neumann's entropy of the fermionic state [23].

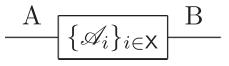
The paper is organized as follows. In Sec. II we give an account of the OPT framework. We set up the basic terminology and we provide the reader with the relevant definitions and assumptions that are necessary in the present work. In Sec. III we give the formal definition of information content and we show that such a definition is well posed. We also prove properties that do not require assumptions on the structure of the theory: Subadditivity and invariance under reversible transformations acting on the state at hand. In Sec. IV we explore the consequence of the steering assumption, namely, the possibility to steer any ensemble of the state by means of one of its dilations. We show that it is possible to assess the reliability of a compression protocol by taking into account the dilations of the state only, instead of considering also all the possible decompositions of them. We generalize the entanglement fidelity and we prove that it can be equivalently used as a figure of merit for defining the information content. In Sec. V we investigate the relation between state purity and the vanishing of the information content. The main result of this section is that all the states with vanishing information content must be necessarily pure. Moreover, we show that the converse is not generally true. Indeed, we first prove that sufficient conditions are the atomicity of parallel composition of states and essential uniqueness of purification. Then, we show that atomicity of parallel composition of states turns out to be a necessary condition for having a null information content on pure states. Thus, any theory violating this property must have pure states with a strictly positive information content. Finally, we show that the accessible information is a lower bound for the information content here introduced. In Sec. VI we show that the information content is simply given by Shannon and von Neumann entropies in classical and quantum information theories, respectively. The point here is to check the collapse of our figure of merit and the one usually used in those theories. We conclude summarizing the results of this work and discussing some open questions in Sec. VII.

II. OPT FRAMEWORK

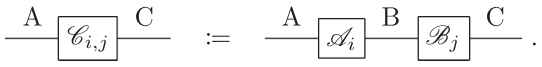
In this section we briefly review the framework of operational probabilistic theories.

A. General description

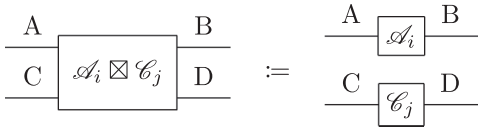
The primitive notions of an operational theory are those of test, event and system. A *test* $\{\mathscr{A}_i\}_{i\in\mathsf{X}}$ is given by a collection of *events*, where i labels the elements of the outcome space X. The *systems* allow for the connection between different tests, and are denoted by 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. 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 $\mathscr{C}_{i,j}$ that are diagrammatically represented by

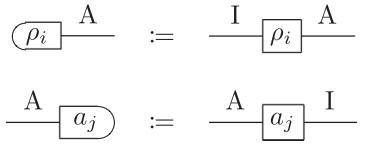


Notice that this definition requires the output system of the events on the left to be necessarily the input system of the events on the right. A *singleton test* is a test whose outcome space set X is a singleton, and the unique event contained in it is called *deterministic*. For any system A there exists a unique *identity* test $\{\mathscr{I}_\mathrm{A}\}$ such that $\mathscr{A}\,\mathscr{I}_\mathrm{A} = \mathscr{A}$ ($\mathscr{I}_\mathrm{A}\mathscr{A} = \mathscr{A}$) for any event $\mathscr{A}$. Another operation that can be performed on tests for defining a new test 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, their parallel composition is the test $\{\mathscr{A}_i \boxtimes \mathscr{C}_j\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$. Diagrammatically



The parallel composition operation commutes with the sequential one, namely, $(\mathscr{E}_h \boxtimes \mathscr{F}_k)(\mathscr{C}_i \boxtimes \mathscr{D}_j) = (\mathscr{E}_h\mathscr{C}_i) \boxtimes (\mathscr{F}_k\mathscr{D}_j)$.

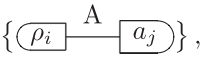
There is a special kind of system, the *trivial system* I, satisfying $\mathrm{AI} = \mathrm{IA} = \mathrm{A}$ for every system A. Tests with I as input system and A as the output one are called *preparation tests* of A, while tests with input system A and I as output are named *observation tests* of 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 test.

Preparation and observation events will also be denoted by using round brackets, respectively, $|\rho)_{\mathrm{A}}$ and ${}_{\mathrm{A}}(a|$, and we will not make explicit the system whenever it is clear from the context.

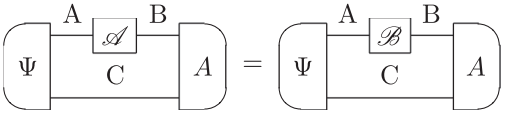
A circuit is a diagram representing an arbitrary test that is obtained by sequential and parallel composition of other tests. We say that a circuit is *closed* when the input and output systems are both the trivial one, namely, when it starts with a preparation test and it ends with an observation test. An *operational probabilistic theory* is an operational theory where any closed circuit (equivalently, any test from the trivial system to istelf) is given by a joint probability distribution conditioned by the tests building the circuit. 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. A simple example is given by a preparation test $\{\rho_i\}_{i\in\mathsf{X}}$ sequentially followed by an observation test $\{a_j\}_{j\in\mathsf{Y}}$:

$$\{p(i,j|\{\rho_i\},\{a_j\})\} := \{$$  $$\},$$

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. 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, $p(i,j) = (a_j|\rho_i)$.

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 $(a|\rho) \neq (a|\sigma)$ for any observation event a. The set of equivalence classes with respect to this relation is called the set of *states* of system A, and it is denoted by $\mathsf{St}(\mathrm{A})$. Similarly, one can define the set of *effects* as the set of equivalence classes of the observation events such that $(a|\rho) \neq (b|\rho)$ for any preparation event ρ, and this is denoted by $\mathsf{Eff}(\mathrm{A})$. The sets of deterministic states and effects will be denoted by $\mathsf{St}_1(\mathrm{A})$ and $\mathsf{Eff}_1(\mathrm{A})$, respectively.

Given the probabilistic structure, states can be seen as functionals on the set of effects and vice versa, 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. The *size* D_{A} of a given system A is simply the dimension of the linear space $\mathsf{St}_{\mathbb{R}}(\mathrm{A})$. 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. Indeed, given $\mathscr{A}$ and $\mathscr{B}$, we say that they are *operationally equivalent* ($\mathscr{A} \sim \mathscr{B}$) if the following identity holds:



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. We then denote the set of all the equivalence classes with $\mathsf{Tr}(\mathrm{A}\to\mathrm{B})$, whose elements are simply called *transformations*. As for states and effects, the set of deterministic transformations will be denoted by $\mathsf{Tr}_1(\mathrm{A}\to\mathrm{B})$. If $\mathscr{U} \in \mathsf{Tr}(\mathrm{A}\to\mathrm{B})$ and there exists $\mathscr{V} \in \mathsf{Tr}(\mathrm{B}\to\mathrm{A})$ such that $\mathscr{V}\mathscr{U} = \mathscr{I}_{\mathrm{A}}$ and $\mathscr{U}\mathscr{V} = \mathscr{I}_{\mathrm{B}}$, we say that $\mathscr{U}$ is *reversible*. Accordingly, two systems are called *operationally equivalent* if there exists a reversible transformation $\mathscr{U} \in \mathsf{Tr}_1(\mathrm{A}\to\mathrm{B})$. A notion that will be useful in the following is that of *asymptotically equivalent systems*.

Definition II.1 (Asymptotical equivalence). Two systems A_1 and A_2 are asymptotically equivalent if

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

(2) There exists a pair of integers $h_1, h_2 < \infty$, $\mathscr{G} \in \mathsf{Tr}_1(\mathrm{A}_2^{\boxtimes h_2} \to \mathrm{A}_1^{\boxtimes h_1})$ and $\mathscr{F} \in \mathsf{Tr}_1(\mathrm{A}_1^{\boxtimes h_1} \to \mathrm{A}_2^{\boxtimes h_2})$ 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:

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

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

Definition II.2 (Refinement of an event). Let $\mathscr{C} \in \mathsf{Tr}(\mathrm{A}\to\mathrm{B})$. A refinement of $\mathscr{C}$ is given by a collection of events $\{\mathscr{D}_i\}_{i\in\mathsf{Y}} \subseteq \mathsf{Tr}(\mathrm{A}\to\mathrm{B})$ such that there exists a test $\{\mathscr{D}_i\}_{i\in\mathsf{X}}$ with $\mathsf{Y}\subseteq\mathsf{X}$ and $\mathscr{C} = \sum_{j\in\mathsf{Y}} \mathscr{D}_j$. 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 $\{\mathscr{D}_i\}_{i\in\mathsf{Y}}$.

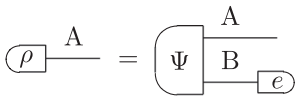
Definition II.3. Given two events $\mathscr{C}, \mathscr{D} \in \mathsf{Tr}(\mathrm{A}\to\mathrm{B})$ we say that $\mathscr{D}$ refines $\mathscr{C}$, and write $\mathscr{D} \prec \mathscr{C}$, if there exist a refinement $\{\mathscr{D}_i\}_{i\in\mathsf{X}}$ of $\mathscr{C}$ such that $\mathscr{D} \in \{\mathscr{D}_i\}_{i\in\mathsf{X}}$

Definition II.4 (Atomic and refinable events). An event $\mathscr{C}$ is called atomic if it admits only trivial refinements. An event is refinable if it is not atomic.

The notion of refinement and refinable events give rise to the definitions of *pure* and *mixed* states.

Definition II.5 (Pure and mixed states). $\rho \in \mathsf{St}(\mathrm{A})$ is called pure if it is atomic, or mixed otherwise. We will denote by $\mathsf{PurSt}(\mathrm{A})$ the set of all the pure states of system A

Definition II.6. 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 $D_\Omega \subseteq D_\rho$

Two special instances of this framework are classical and quantum theory. In classical theory, the systems are associated with real vector spaces $\mathbb{R}^{d_A}$, and different systems are associated with different values of d_A. The set of states is made of substochastic vectors in these spaces, namely, by vectors

$\mathbf{x}$ satisfying $\|\mathbf{x}\|_1 := \sum_{i=1}^{d_A} |x_i| \leqslant 1$, and therefore it is a simplex. The pure states are then represented by the canonical basis vectors $\mathbf{e}_i$ of $\mathbb{R}^{d_A}$. The convex set of effects is given by unit-dominated positive vectors, i.e., those $\mathbf{x} \in \mathbb{R}^{d_A}$ such that $0 \leqslant x_i \leqslant 1$ for any $i = 0, \ldots, d_A$. Transformations from system A to system B are represented by $d^B \times d^A$ substochastic matrices $\mathbf{M}$ acting on the probability vectors by multiplication $\mathbf{x} \to \mathbf{M}\mathbf{x}$. Recall that a matrix is substochastic when each column is a substochastic vector. Sequential and parallel composition are trivially given by matrix multiplication and tensor product, respectively.

In quantum theory systems are associated with Hilbert spaces, and different systems are represented by spaces of different dimension d (we will assume $d < \infty$ whenever we will refer to results relative to quantum theory). The convex set of states of a system A is given by subnormalized density matrices ρ on the associated space $\mathcal{H}_A$, i.e., matrices such that $\rho > 0$ and $\mathrm{Tr}(\rho) \leqslant 1$. The convex set of effects is given by functionals (acting on the set of states) of the form $\mathrm{Tr}(\cdot E)$, where E is a positive matrix dominated by the identity, $0 \leqslant E \leqslant I$. Transformations in $\mathsf{Tr}(\mathrm{A} \to \mathrm{B})$ are mathematically represented by completely positive trace nonincreasing maps. These have a Kraus decomposition, namely, if $\mathscr{E} \in \mathsf{Tr}(\mathrm{A} \to \mathrm{B})$ there exists a set of operators $\{E_i\} \subseteq \mathcal{L}(\mathcal{H}_B, \mathcal{H}_A)$ such that $\mathscr{E}(\cdot) = \sum_i E_i^\dagger \cdot E_i$. Sequential composition is simply given by composition of maps, and the parallel one is represented by the tensor product, as in the classical case.

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 [11].

Definition II.7 (Operational norm). The norm of an element $\rho \in \mathsf{St}(\mathrm{A})_{\mathbb{R}}$ is defined as

$$\|\rho\|_{\mathrm{op}} := \sup_{a \in \mathsf{Eff}(\mathrm{A})} (2a - e_A|\rho),$$

where e_A is the deterministic effect obtained by the coarse graining of the observation test containing a.

This norm satisfies a monotonicity property, as stated in the following lemma.

Lemma II.1 (Monotonicity of the operational norm). For any $\delta \in \mathsf{St}_{\mathbb{R}}(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Tr}(\mathrm{A}, \mathrm{B})$ the following inequality holds:

$$\|\mathscr{C}\delta\|_{\mathrm{op}} \leqslant \|\delta\|_{\mathrm{op}}, \tag{2}$$

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

This notion of norm, which is valid for any OPT, reduces to the trace norm in quantum theory.

B. Restricting the class of theories

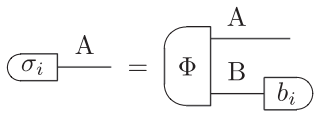
Upon marginalization over the observation test, one can define the preparation probability conditioned by the test $\{a_j\}_{j \in \mathsf{Y}}$ as $p(\rho_i|\{a_j\}) := \sum_j p(i, j)$. Generally, the preparation probability is not one, unless the preparation test $\{\rho_i\}_{i \in \mathsf{X}}$ is the singleton, i.e., the state is deterministic. Moreover, as it is clear by its definition, it can also depend on the observation test we are marginalizing over. Usually, the causality condition is expressed as a *no signaling from the future* principle, namely, by saying that preparation probabilities are actually independent of the chosen observation test, which is equivalent to state the uniqueness of the deterministic effect. In this paper we will adopt a stronger form of causality, that is the following one.

Assumption 1 (Causal theories). An OPT satisfies strong causality if for every test $\{\mathscr{A}_i\}_{i \in \mathsf{X}}$ and every collection of tests $\{\mathscr{B}_j^i\}_{j \in \mathsf{Y}}$ labeled by $j \in \mathsf{Y}$, the collection of events $\{\mathscr{C}_{i,j}\}_{(i,j) \in \mathsf{X} \times \mathsf{Y}}$ with

is a test of the theory.

One can show that the above statement implies uniqueness of the deterministic effect [10,11].

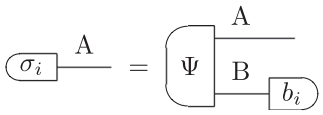
Another assumption that we will use in some sections of this work stems from the steering property of quantum theory. This asserts that, given a state $\rho \in \mathsf{St}(\mathrm{A})$ and a purification $\Phi \in \mathsf{PurSt}(\mathrm{AB})$ of ρ, for any decomposition $\sum_{i \in \mathsf{X}} p_i \sigma_i$ of ρ there exists an observation test $\{b_i\}_{i \in \mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{B})$ such that



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

This feature cannot be assumed as it stands: The first reason is that a generic OPT may not encompass the existence of a purification for any state of the theory (and classical theory is a trivial example), therefore we are led to consider dilations instead of purifications. Second, there is no reason why one should be able to steer any decomposition by means of the same dilation. Thus, for a generic theory, one can state the steering as follows.

Assumption 2 (Steering). 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 steering assumption must be a dilation of ρ, as one can easily verify upon summing over $i \in \mathsf{X}$. The stronger steering feature satisfied by quantum theory can actually be proven to hold in any OPT satisfying atomicity of parallel composition of states, existence and uniqueness (up to reversible channels) of purification and perfect discriminability as axioms. For the present purposes we choose to state steering as a property that an OPT may or may not satisfy rather than discussing the conditions under which it holds.

Another property that we will assume throughout the paper is digitizability.

Assumption 3 (Digitizability). 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 $\mathscr{C} \in \mathsf{Tr}_1(\mathrm{X} \to \mathrm{B}^{\boxtimes k})$ and $\mathscr{F} \in \mathsf{Tr}_1(\mathrm{B}^{\boxtimes k} \to \mathrm{X})$ such that $\mathscr{F} \circ \mathscr{C} = \mathscr{I}_X$. Moreover, if B_1 and B_2 are two such systems, then they are asymptotically equivalent.

The above assumption holds in quantum theory, since any qudit system can be taken as elementary. Moreover, let us consider two different qudits with dimension d_1 and d_2, respectively. Generally, the equation $d_1^N = d_2^M$ may have no integer solutions (for instance, when both d_1 and d_2 are prime). However, the smallest integer M such that we can isometrically embed $\mathcal{H}_1^{\otimes N}$ into $\mathcal{H}_2^{\otimes M}$ is given by $\lceil N \log_{d_2} d_1 \rceil$ which is such that $\lim_{N\to\infty} \frac{\lceil N \log_{d_2} d_1 \rceil}{N} = \log_{d_2} d_1$. Similarly, $\lceil M \log_{d_1} d_2 \rceil$ is needed for an isometric embedding of $\mathcal{H}_2^{\otimes M}$ into $\mathcal{H}_1^{\otimes N}$, and $\lim_{N\to\infty} \frac{\lceil M \log_{d_1} d_2 \rceil}{M} = \log_{d_1} d_2 (= 1/\log_{d_2} d_1)$. In any theory satisfying the assumption of digitizability we can always encode the state of our system on the parallel composition of a sufficiently large number of elementary systems, which we can think of as a generalization of the qubit system for quantum theory. The request of digitizability comes from the need of a unit for the amount of information required for storing 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.

We want to stress the fact that the assumption of digitizability 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 nondigitizable theory should contain infinitely many inequivalent system types, even asymptotically (see Definition II.1), and this immediately brings us into an unexplored territory of wild theories.

The nonlocal boxes [21] provide us with a non trivial example of an OPT satisfying the digitizability 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 [16,24]—nonlocal boxes are presented in terms of the geometry of their state space, and focusing on the correlations that measurements can produce, disregarding the behavior 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. [16]. Nonlocal 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 [25]) opening the route to a generalization of this property in the infinite case.

III. INFORMATION CONTENT IN OPT

Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and consider N copies of the system on which we have prepared the same state ρ and let M be a positive integer. A compression scheme is then a pair of maps $\mathscr{E} \in \mathsf{Tr}_1(\mathrm{A}^{\boxtimes N} \to \mathrm{B}^{\boxtimes M})$, $\mathscr{D} \in \mathsf{Tr}_1(\mathrm{B}^{\boxtimes M} \to \mathrm{A}^{\boxtimes N})$. Sometimes we will denote by $\mathscr{C}$ the composition $\mathscr{D}\mathscr{E}$.

Definition III.1. An (ε, N)-reliable compression scheme $(\mathscr{E}, \mathscr{D})$ is such that

$$\sup_{\mathrm{C},\{\Psi_i\}} \sum_{i\in\mathsf{X}} \|[(\mathscr{D}\mathscr{E}) \boxtimes \mathscr{I}_{\mathrm{C}}]\Psi_i - \Psi_i\| < \varepsilon,$$

where $\{\Psi_i\}_{i\in\mathsf{X}} \subseteq \mathsf{St}_1(\mathrm{A}^{\boxtimes N}\mathrm{C})$ denotes a refinement of any dilation of $\rho^{\boxtimes N}$. For fixed N, M we denote with $E_{N,M,\varepsilon}(\rho)$ the set of ε-reliable compression schemes.

Definition III.2 (Information Content). Let $\rho \in \mathsf{St}_1(\mathrm{A})$. We define the *smallest achievable compression ratio* for length N to tolerance ε as follows:

$$R_{N,\varepsilon}(\rho) := \frac{\min\{M : E_{N,M,\varepsilon}(\rho) \neq \emptyset\}}{N}. \tag{3}$$

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

$$I(\rho) := \lim_{\varepsilon\to 0} \limsup_{N\to\infty} R_{N,\varepsilon}(\rho). \tag{4}$$

Proposition III.1. $I(\rho)$ is well defined for every $\rho \in \mathsf{St}(\mathrm{A})$ and every system A.

Proof. First, 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 $\mathscr{E} \in \mathsf{Tr}_1(\mathrm{A}^{\boxtimes N} \to \mathrm{B}^{\boxtimes K})$, $\mathscr{D} \in \mathsf{Tr}_1(\mathrm{B}^{\boxtimes K} \to \mathrm{A}^{\boxtimes N})$ 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. (3) is always finite. Moreover, it is immediate to realize 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. (3) 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 $\varepsilon \leqslant \varepsilon'$, 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, fix N, ε, let $\rho \in \mathsf{St}_1(\mathrm{A})$, and let

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

be the minimum number of obits B_1 and B_2 needed for an ε-optimal encoding, respectively. Rephrasing the first equation in (1), 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 II.1 from $M_{1,N}$ to $M_2^{\min}(M_{1,N})$ (see item 3 in Definition II.1), we have an ε-optimal encoding of $\rho^{\boxtimes N}$ onto $M_2^{\min}(M_{1,N})$ obits B_2, therefore $M_{2,N} \leqslant M_2^{\min}(M_{1,N})$ and this implies

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

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

Proposition III.2 (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:

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

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 $\overline{K}:=\min\{M: E_{N,M,\varepsilon}(\rho)\neq\emptyset\}$ 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 $\mathscr{C}^\rho:=\mathscr{D}^\rho\mathscr{E}^\rho$. 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:

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

where we used the triangle inequality for the operational norm. Thus $E_{N,\overline{K}+\overline{L},2\varepsilon}(\Psi)\neq\emptyset$ and this implies that

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

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

We notice that, in order to compute the information content, one can test the compression schemes on pure decompositions $\{p_i,\Phi_i\}$ only. More precisely, let $E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho)$ be the set of schemes which are (ε,N)-reliable according to the following criterion:

$$\sup\left(\sum_i p_i\|(\mathscr{D}\mathscr{E}\boxtimes\mathscr{I})(\Phi_i)-\Phi_i\|_{\mathrm{op}}\right)<\varepsilon,\qquad(5)$$

where the supremum is taken on all the pure decompositions $\{p_i,\Phi_i\}$ of any $\Omega\in D_{\rho^{\boxtimes N}}$. Let $I^{\mathrm{pur}}(\rho)$ be the information content computed restricting to such maps:

$$\begin{aligned}R^{\mathrm{pur}}_{N,\varepsilon}(\rho)&:=\frac{\min\left\{M: E^{\mathrm{pur}}_{N,M,\varepsilon}(\rho)\neq\emptyset\right\}}{N}\\I^{\mathrm{pur}}(\rho)&:=\lim_{\varepsilon\to 0}\limsup_{N\to\infty}R^{\mathrm{pur}}_{N,\varepsilon}(\rho).\end{aligned}$$

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

Lemma III.1. Let $\rho\in\mathsf{St}_1(\mathrm{A})$, then $I(\rho)=I^{\mathrm{pur}}(\rho)$

Proof. On the one hand, we trivially have $I^{\mathrm{pur}}(\rho)\leqslant I(\rho)$. On the other hand, let $\{\Psi_i\}_{i\in\mathsf{X}}$ be a refinement of $\Omega\in D_{\rho^{\boxtimes N}}$. For any i we can further decompose Ψ_i in terms of pure states $\{q^i_j,\Phi_{i,j}\}_{j\in\mathsf{Y}}$, with $\sum_{i,j}q^i_j=1$. Therefore $\{q^i_j,\Phi_{i,j}\}_{(i,j)\in\mathsf{X}\times\mathsf{Y}}$ is a pure state decomposition of Ω, and by the triangle inequality one has

$$\begin{aligned}&\sum_{i\in\mathsf{X}}\|[(\mathscr{D}\circ\mathscr{C})\boxtimes\mathscr{I}]\Psi_i-\Psi_i\|_{\mathrm{op}}\\&\quad\leqslant\sum_{(i,j)\in\mathsf{X}\times\mathsf{Y}}q^i_j\|[(\mathscr{D}\circ\mathscr{C})\boxtimes\mathscr{I}]\Phi_{i,j}-\Phi_{i,j}\|_{\mathrm{op}}.\end{aligned}$$

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

Proposition III.3. Let $\rho\in\mathsf{St}_1(\mathrm{A})$ and $\mathscr{U}\in\mathsf{Tr}_1(\mathrm{A})$ be a reversible channel, then $I(\rho)=I(\mathscr{U}(\rho))$.

Proof. We show that $E_{N,M,\varepsilon}(\mathscr{U}(\rho))\neq\emptyset\Rightarrow E_{N,M,\varepsilon}(\rho)\neq\emptyset$. 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 $R_{N,\varepsilon}(\rho)\leqslant R_{N,\varepsilon}(\mathscr{U}(\rho))$, and then $I(\rho)\leqslant I[\mathscr{U}(\rho)]$. The reverse inequality is now trivial

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

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

IV. STEERING: INFORMATION CONTENT FROM DILATIONS

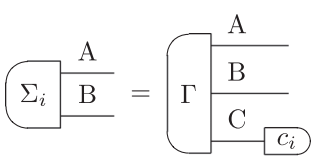
In this section we show that in an OPT satisfying the steering property (assumption 2) the information content of a state can be computed by considering the action of the compression schemes on the set $D_{\rho^{\boxtimes N}}$ only.

Lemma IV.1. Let $\rho\in\mathsf{St}_1(\mathrm{A})$, $\mathscr{C}\in\mathsf{Tr}_1(\mathrm{A})$ and $\varepsilon>0$. If $\|\mathscr{C}\boxtimes\mathscr{I}(\Psi)-\Psi\|<\varepsilon$ for any $\Psi\in D_\rho$ and Assumption 2 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 Assumption 2 there exists $\Gamma\in D_\Omega\subseteq D_\rho$ and an observation test $\{c_i\}_{i\in\mathsf{X}}$ such that



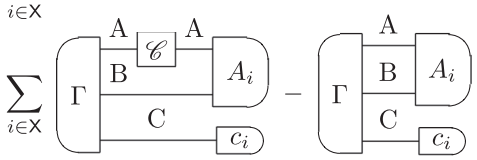
$$,\qquad\forall i\in\mathsf{X}.$$

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 (see Appendix A), 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 IV.1. Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and consider

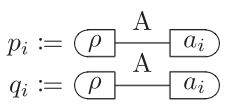
$$I^{\mathrm{dil}}(\rho) := \lim_{\varepsilon\to 0} \limsup_{N\to\infty} \frac{\min\left\{M : E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho) \neq \emptyset\right\}}{N}, \tag{6}$$

where $E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$ is the set of the compression schemes $(\mathscr{E}, \mathscr{D})$ such that $\sup_{\Psi\in D_{\rho^{\boxtimes N}}} \|\mathscr{C} \boxtimes \mathscr{I}(\Psi) - \Psi\| < \varepsilon$. Then $I(\rho) = I^{\mathrm{dil}}(\rho)$.

Proof. This is a straightforward consequence of the lemma. 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 $I^{\mathrm{dil}}(\rho) \leqslant I(\rho)$. The inclusion $E_{N,M,\varepsilon}(\rho) \supseteq E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$ follows by Lemma IV.1, and this implies the thesis. ■

We now prove some bounds that involve a quantity generalizing the classical and quantum fidelity. For this purpose, we consider a definition of fidelity [19] that can be adopted in the OPT framework, which reduces to the classical or quantum one in the in the respective theories. ■

Definition IV.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 $\mathbf{p} := p_i$ and $\mathbf{q} := q_i$ the probability distributions defined by



Then one can define the fidelity between ρ and σ as

$$F(\rho, \sigma) := \inf_{\{a_i\}_{i\in\mathsf{X}}} F_c(\mathbf{p}, \mathbf{q}), \tag{7}$$

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 $F(\rho, \sigma) \leqslant 1$, with equality if and only if $\rho = \sigma$. Fidelity satisfies the following property, that generalizes the Fuchs–van de Graaf inequality [26] and is relevant for the present work.

Proposition IV.2. Let $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$. The following inequalities hold:

$$1 - F(\rho, \sigma) \leqslant \tfrac{1}{2}\|\rho - \sigma\|_{\mathrm{op}} \leqslant \sqrt{1 - F(\rho, \sigma)^2}. \tag{8}$$

Proposition IV.3 (Monotonicity). Let $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Tr}_1(\mathrm{A}, \mathrm{B})$. The following inequality holds:

$$F(\mathscr{C}(\rho), \mathscr{C}(\sigma)) \geqslant F(\rho, \sigma). \tag{9}$$

In quantum theory we have a notion, the so called *entanglement fidelity*, which measures how well correlations with an environment are preserved by a given channel acting on our local system. If $\rho \in \mathsf{St}_1(\mathrm{A})$ is the state of our local system, $\Phi \in \mathsf{PurSt}(\mathrm{AB})$ is a purification of ρ and $\mathscr{C} \in \mathsf{Tr}_1(\mathrm{A} \to \mathrm{C})$ the channel locally applied to A, then the entanglement fidelity is defined as the square of the Uhlmann one between input and output, $F[\Phi, \mathscr{C} \otimes \mathscr{I}(\Phi)]^2$. This is a well defined quantity since it is independent of the chosen purification. By means of the generalized notion of fidelity in (7) we can define an analogous of the entanglement fidelity in the OPT framework. Again, we must be aware of the fact that in a generic OPT, there may be states that cannot be purified (mixed states in classical theories are a rather trivial example). In order to encompass the most general situation, we refer to dilations rather than focusing on purifications. Moreover, we want to define a quantity that is independent of the particular dilation, and we are thus led to the following definition.

Definition IV.2. Let $\rho \in \mathsf{St}(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Tr}_1(\mathrm{A} \to \mathrm{C})$. We define the correlation fidelity as follows:

$$F(\rho, \mathscr{C}) = \inf_{\Psi\in D_\rho} F[\Psi, \mathscr{C} \boxtimes \mathscr{I}(\Psi)]^2. \tag{10}$$

By means of the generalized Fuchs–van de Graaf inequality (8) we can see that the correlation fidelity can be used as an equivalent figure of merit on $D_{\rho^{\boxtimes N}}$. More precisely, the following proposition holds.

Proposition IV.4. Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and define

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

where

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

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

Proof. This is a straightforward consequence of Proposition IV.2. Let $(\mathscr{E}, \mathscr{D}) \in E^{\mathrm{dil}}_{N,M,\varepsilon}(\rho)$. By the first inequality in (8) we have that

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

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

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 IV.2 we have that

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

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 $I^F(\rho) \geqslant I^{\mathrm{dil}}(\rho)$ follows. ■

V. INFORMATION CONTENT AND STATE PURITY

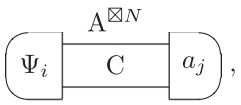
In the following we will assume that

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

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

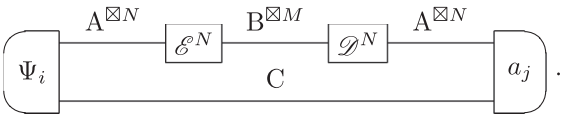
$$\sum_i e_{\mathrm{C}} \circ \Psi_i = \rho^{\boxtimes N}.$$

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

$p_{i,j} :=$ 

, (11)

and

$q_{i,j} :=$ 

. (12)

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 $P(X = x_i, \tilde{Y} = y_j) = q_{i,j}$:

$$I(X : Y) := \sum_{i,j} p_{i,j} \log_2 \frac{p_{i,j}}{p_i^X p_j^Y},$$

$$I(X : \tilde{Y}) := \sum_{i,j} q_{i,j} \log_2 \frac{q_{i,j}}{q_i^X q_j^{\tilde{Y}}},$$

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

Definition V.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

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

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:

$$R^C_{\delta,N}(\rho) := \frac{\min\left\{M \mid E^C_{N,M,\delta}(\rho) \neq \emptyset\right\}}{N}, \tag{13}$$

$$R^C_{\delta}(\rho) := \limsup_{N\to\infty} R^C_{\delta,N}(\rho), \tag{14}$$

$$I^C(\rho) := \lim_{\delta\to 0} R^C_{\delta}(\rho). \tag{15}$$

The last quantity above satisfies the following lemmas.

Lemma V.1. Let $\rho \in \mathsf{St}_1(\mathrm{A})$. Then $I(\rho) \geqslant I^C(\rho)$.

The proof can be found in Appendix B.

In proving the following 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 4 (Regular scaling). For every type of system A, there exists a constant $k_{\mathrm{A}} > 0$ such that the size $D(N) := D_{\mathrm{A}^{\boxtimes N}}$ of the compound system $\mathrm{A}^{\boxtimes N}$ satisfies $D(N) \leqslant k_{\mathrm{A}} D(1)^N$.

Lemma V.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 C.

Proposition V.1. Let $\rho \in \mathsf{St}_1(\mathrm{A})$. If $I(\rho) = 0$ then ρ is a pure state.

Proof. Let ρ be mixed. By Lemmas V.1 and V.2

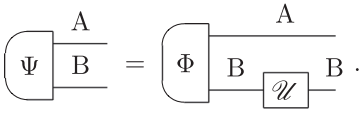
$$I(\rho) \geqslant I^C(\rho) > 0,$$

whence the thesis. ∎

We now use the above results to prove some general facts about theories with *essentially unique purification* and *atomicity of parallel composition* for states. For this purpose, let us start reminding in the first place the definition of the latter requirements.

Definition V.2 (Existence of purification). We say that an OPT satisfies purification if for any $\rho \in \mathsf{St}(\mathrm{A})$ one has $P_\rho \neq \emptyset$.

Definition V.3 (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



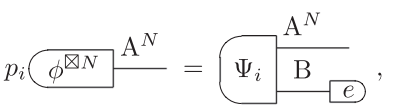
. (16)

Definition V.4 (Atomicity of parallel composition of states). We say that an OPT satisfies atomicity of parallel composition of states if for any pair $\phi \in \mathsf{PurSt}(\mathrm{A})$ and $\psi \in \mathsf{PurSt}(\mathrm{B})$ we also have $\phi \boxtimes \psi \in \mathsf{PurSt}(\mathrm{AB})$.

We now prove that, in every strongly causal theory that satisfies regular scaling and uniqueness of purification, null information content of pure states is equivalent to atomicity of parallel composition of states. We stress that the requirement of *existence* of purification is not needed, but only its uniqueness. In other words, *if* a state has a purification, then the latter is unique, even though it needs not have one. The following result is particularly interesting because it provides an alternative way of understanding the operational content of atomicity of parallel composition.

Proposition V.2. Let us consider strongly causal OPT satisfying regular scaling. Then the requirements of essential uniqueness of purification and atomicity 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 $\phi \in \mathsf{PurSt}(\mathrm{A})$, atomicity of parallel composition of states holds.

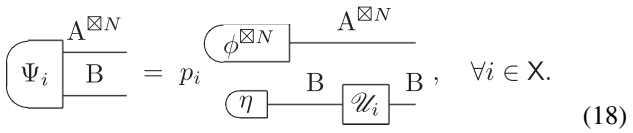
Proof. By Lemma III.1 we have $I(\phi) = I^{\mathrm{pur}}(\phi)$. Now, let us fix N and consider a dilation Ω of $\phi^{\boxtimes N}$. Let $\{\Psi_i\}_{i\in\mathsf{X}}$ be a pure decomposition of Ω, then by purity of $\phi^{\boxtimes N}$ we must have



, $\quad \forall i \in \mathsf{X}.$ (17)

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

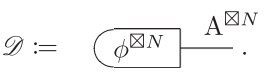
uniqueness of purifications



(18)

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:



(19)

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 $E^{\rm pur}_{N,0,\varepsilon}(\rho) \neq \emptyset$ and then $I(\rho) = I^{\rm pur}(\rho) = 0$.

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 III.2 we have that

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

Thus $I(\rho \boxtimes \sigma) = 0$, and by proposition V.1 $\rho \boxtimes \sigma$ is pure, namely, Assumption V.4 holds. ■

It is interesting to observe that, due to the above proposition, one can exhibit operational probabilistic theories having pure states with nonvanishing information content. In Ref. [27] the authors construct a theory, *bilocal classical theory*, where all systems are classical (the set of states is a simplex), but with a parallel composition rule that differs from the one of classical information theory, thereby violating Assumption V.4. Accordingly, bilocal classical theory must have pure states with nonnull information content.

VI. INFORMATION CONTENT IN QUANTUM AND CLASSICAL INFORMATION THEORY

Before restricting to the quantum case, we prove the following lemma concerning the correlation fidelity defined in Definition IV.2.

Lemma VI.1. Let $\rho \in \mathsf{St}_1(\mathrm{A})$ and $\mathscr{C} \in \mathsf{Tr}_1(\mathrm{A})$. If every state has a purification (Definition V.2), then one has

$$F(\rho, \mathscr{C}) = \inf_{\Phi \in P_\rho} F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2. \tag{20}$$

Moreover, in an OPT with essential uniqueness of purification (Definition V.4) and atomicity of parallel composition of states (Definition V.3), for any $\Phi \in P_\rho$ one has

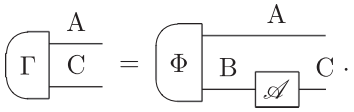
$$F(\rho, \mathscr{C}) = F[\Phi, \mathscr{C} \boxtimes \mathscr{I}(\Phi)]^2. \tag{21}$$

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 (Proposition IV.3) we have

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

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

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 hold (see Definitions V.4 and V.3), then, for any fixed purification Φ in P_ρ, and any other Γ in P_ρ, there exists a channel $\mathscr{A} \in \mathsf{Tr}_1(\mathrm{B}, \mathrm{C})$ such that



(22)

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

We recall the statement of the Schumacher theorem. Let $\rho \in \mathsf{St}_1(\rho)$ and $(\mathscr{E}, \mathscr{D})$ be a compression scheme

Theorem VI.1 (Schumacher). Let $\rho \in \mathsf{St}_1(\mathrm{A})$ with $\mathcal{H}_\mathrm{A}$ the Hilbert space corresponding to the quantum system A, let $(\mathscr{E}, \mathscr{D})$ be a compression scheme and define its ratio R as

$$R := \frac{\log[\dim[\mathsf{Supp}(\mathscr{E}(\rho^{\otimes N}))]]}{N}.$$

For every $\varepsilon > 0$ and $R > S(\rho)$ there exists N_0 such that $\forall N \geqslant N_0$ there exists a compression scheme with ratio R such that $F(\rho^{\otimes N}, \mathscr{D}\mathscr{E}) > 1 - \varepsilon$. Conversely, for every $R < S(\rho)$ there is $\varepsilon > 0$ such that for every compression scheme $(\mathscr{E}, \mathscr{D})$ with ratio R one has $F(\rho^{\otimes N}, \mathscr{D}\mathscr{E}) \leqslant \varepsilon$.

Proposition VI.1. 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 $I^F(\rho) \leqslant S(\rho)$. Let $\delta > 0$, $R \in (S(\rho), S(\rho) + \delta]$ and $\varepsilon > 0$. By the direct part of the Schumacher theorem there exists a N_0 such that for any $N \geqslant N_0$ there is a (N, ε)-reliable compression scheme with rate R. Thus, upon embedding $\mathsf{Supp}[\mathscr{E}(\rho^{\otimes N})]$ in $\lceil NR \rceil$ qubits, by using a suitable isometry, we have $E^F_{N,\lceil NR \rceil,\varepsilon}(\rho) \neq \emptyset$ for any $N \geqslant N_0$. This implies

$$\begin{aligned} \limsup_{N\to\infty} \frac{\min\left\{M : E^F_{N,M,\varepsilon}(\rho) \neq \emptyset\right\}}{N} \\ \leqslant \lim_{N\to\infty} \frac{\lceil NR \rceil}{N} = R \leqslant S(\rho) + \delta. \end{aligned}$$

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

Now let $\delta > 0$ and consider $M/\overline{N}$ such that $S(\rho) - \delta \leqslant M/\overline{N} < S(\rho)$. By the converse part of Schumacher theorem there exists $\overline{\varepsilon} > 0$ such that for any compression scheme with ratio $M/\overline{N}$ one has $F(\rho^{\otimes N}, \mathscr{D}\mathscr{E}) \leqslant 1 - \overline{\varepsilon}$. 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

$$\begin{aligned} S(\rho) - \delta \leqslant \frac{M}{\overline{N}} < \frac{\min\left\{L : E^F_{k\overline{N},L,\overline{\varepsilon}}(\rho) \neq \emptyset\right\}}{k\overline{N}} \leqslant \\ \leqslant \frac{\min\left\{L : E^F_{k\overline{N},L,\varepsilon}(\rho) \neq \emptyset\right\}}{k\overline{N}}. \end{aligned}$$

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

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

for any $0<\varepsilon<\overline{\varepsilon}$. By taking the $\lim_{\varepsilon\to 0}$ and the arbitrariness of δ we finally get $S(\rho)\leqslant I^F(\rho)$. The statement then follows by the fact that in quantum theory one has $I^F(\rho)=I(\rho)$ (Propositions IV.1 and IV.4). ■

Let us now turn our focus to the classical case. In this setting, the input and the output of the compression scheme are given by strings of N letters drawn from an alphabet χ. 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 $p_{i_1,\dots,i_N}=p_{i_1}\dots p_{i_N}$. The probability $p(e)$ of emitting an output string which is different form the input one is often considered in the literature as a figure of merit. More formally, let C be the Markov matrix representing the composition of the compression and decompression maps, and $\mathbf{i}:=i_1\dots i_N$ define the input string, then the error probability is defined as

$$p^C(e):=\sum_{\mathbf{i}}\sum_{\mathbf{j}\neq\mathbf{i}} p(\mathbf{i}\neq\mathbf{j}|\mathbf{i}) = 1-\sum_{\mathbf{i}} C_{\mathbf{i},\mathbf{i}}p_{\mathbf{i}}.$$

The set of states of classical theory 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 in the ith position: $(e_i)_j=\delta_{i,j}$. Since C is a stochastic matrix we have the following chain of equalities:

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

namely,

$$p^C(e)=\sum_{\mathbf{i}} p_{\mathbf{i}}\tfrac{1}{2}\|C\mathbf{e}_{\mathbf{i}}-\mathbf{e}_{\mathbf{i}}\|_1.$$

Now consider *the unique* pure decomposition of some dilation $\mathbf{\Pi}\in D_{\mathbf{p}^{\otimes N}}$. This is given by $\mathbf{\Pi}=\sum_{\mathbf{i},j}\Pi_{ij}\mathbf{e}_{\mathbf{i}}\otimes\mathbf{e}_j$ 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

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

having used the fact that for any j

$$\|(C-I)\mathbf{e}_{\mathbf{i}}\otimes\mathbf{e}_j\|_1=\|(C-I)\mathbf{e}_{\mathbf{i}}\|_1.$$

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

$$p^C(e)=\frac{1}{2}\sum_{\mathbf{i},j}\Pi_{\mathbf{i}j}\|(C\otimes I)\mathbf{e}_{\mathbf{i}}\otimes\mathbf{e}_j-\mathbf{e}_{\mathbf{i}}\otimes\mathbf{e}_j\|_1.$$

Namely, in the classical case, the error probability is exactly our figure of merit in (5).

Now, one can use the first Shannon theorem in order to prove that the information content of a classical state is exactly its Shannon entropy

Theorem VI.2 (Shannon). Let $\mathbf{p}\in\mathsf{St}_1(\mathrm{A})$ be a classical state, let $(\mathscr{E},\mathscr{D})$ be a compression scheme and define its ratio R as

$$R:=\frac{\log_2|\mathscr{E}(\mathsf{Rng}(\mathrm{X}^N))|}{N}.$$

For every $\varepsilon>0$ and $R>H(X)$ there exists N_0 such that $\forall N\geqslant N_0$ there exists a compression scheme with ratio R such that $p^C(e)<\varepsilon$. Conversely, for every $R<H(X)$ there is $\varepsilon>0$ such that for every compression scheme $(\mathscr{E},\mathscr{D})$ with ratio R one has $p^C(e)\geqslant\varepsilon$.

A proposition analogous to Proposition VI.1 holds for the classical case, whose proof is essentially the same. The main issue in both cases is to correctly identify the figure of merit that must be adopted in order to define the information content.

VII. CONCLUSION

We have defined the information content for a source of information of an arbitrary operational probabilistic theory. The only assumption needed is that of digitizability: A theory is digitizable 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 a 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: (1) any preparation of ensembles that average to the considered state must be indistinguishable from leaving the preparation untouched and (2) 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, in the hypothesis of steering of ensembles, we show that the information content can be computed using simple figures of merit, e.g., a generalization 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. In quantum theory both the entanglement fidelity and the average input-output fidelity for an arbitrary decomposition of the state representing the source identify the von Neumann entropy as the minimal compression rate. This opens a relevant question: Which are the minimal assumptions behind the collapse of "global" and "local" figures of merit, namely, quantifiers of the ability to recover the correlations of the source, and the local preparations of the source, respectively?

Like Shannon's and von Neumann's entropy, we proved that the information content is subadditive, and can be used to measure the purity of a state. Indeed both Shannon's and von Neumann's entropy vanish if and only if the state is pure, and here we show that the information content has this feature as well. 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 the light of the above results we propose the information content as a candidate entropic quantity generalizing Shannon entropy of classical systems and von Neumann entropy of quantum systems. A basic message across the literature on general probabilistic theories [17–19] is that a theory is usually not *monoentropic*, namely, multiple entropic quantities can be defined, each one reducing to Shannon's and von Neumann's entropy in classical and quantum theory, respectively. The notions of entropy usually considered are defined in terms of classical information quantities as follows: (1) 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. (2) 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. (3) 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*.

For a general theory the above quantities can be very different and violate some of the typical features of Shannon and von Neumann entropies, such as *concavity* and *strong subadditivity*. It is known [17–19] that measurement entropy is both subadditive and concave but in general it is not strongly subadditive and does not provide a measure of purity, while decomposition entropy is generally neither subadditive nor concave. For example, in Ref. [17] it is shown that, for nonlocal boxes, the decomposition entropy is not concave. Less is known about the third entropic quantity given in terms of the Shannon mutual information, which still could satisfy all features of Shannon and von Neumann entropies.

One of the main outcome of this manuscript are a series of results that can be used to clarify which of the possible entropies of a general probabilistic theory has operational meaning in terms of optimal compression ratio. We started here the analysis of the relation between information content and other entropies of a general probabilistic theory focusing on the accessible information. On one hand, both quantities provide a measure of purity of a state, and on the other hand we proved that the accessible information is a lower bound for information content. An important open question is under what conditions the information content, which by definition is the optimal compression ratio, coincides with the accessible information in general.

Finally, we leave the question open as whether nondigitizable theories exist, or one can figure out a counterexample. While we conjecture that the second choice is the case, it is very hard to exhibit a nondigitizable theory, precisely because it is hard to conceive a pair of systems that are not asymptotically equivalent, though it is intuitive that, e.g., a system with a state space that is a polyhedron cannot be asymptotically equivalent to a system whose state space is an ellipsoid. A nondigitizable theory should contain infinitely many system types, all pairwise asymptotically inequivalent.

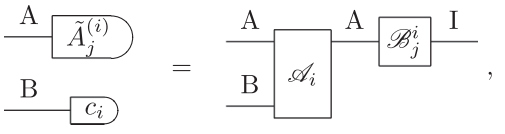
ACKNOWLEDGMENTS

A.T. acknowledges financial support from the Elvia and Federico Faggin Foundation through Silicon Valley Community Foundation, Grant No. 2020-214365.

APPENDIX A: A SIMPLE LEMMA

Lemma A.1. Let $\{c\}_{i\in\mathsf{X}} \subseteq \mathsf{Eff}(\mathrm{B})$ 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 causality. 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



(A1)

with $\mathscr{A}_i := \mathscr{I}_\mathrm{A} \boxtimes c_i$ and $\mathscr{B}_j^i := \tilde{A}_j^{(i)}$. Therefore, 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. ■

APPENDIX B: PROOF OF LEMMA V.1

We start by defining the following number:

$$\zeta(N,\delta) := \sup\left\{\varepsilon \mid E_{N,M,\varepsilon}(\rho) \subseteq E^C_{N,M,\delta}(\rho)\right\}. \tag{B1}$$

First, 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 $(\mathscr{E},\mathscr{D}) \in E_{N,M,\zeta(N,\delta)}$. By definition of $E_{N,M,\zeta(N,\delta)}$ 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

$$\limsup_{N\to\infty} R_{\zeta(N,\delta),N}(\rho) \geqslant R^C_\delta(\rho).$$

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 $\liminf_{N\to\infty}\zeta(N,\delta) =: \bar{\zeta}(\delta) > 0$.

Let us start analyzing case 2. In this case, by definition of limit inferior, one has

$$\forall\delta,\gamma > 0$$
$$\begin{cases}\exists N_0, & \forall N \geqslant N_0, \quad \zeta(N,\delta) > \bar{\zeta}(\delta) - \gamma,\\ \forall N_0, & \exists N \geqslant N_0, \quad \zeta(N,\delta) < \bar{\zeta}(\delta) + \gamma.\end{cases}$$

This implies that for every $\delta > 0$ and every positive γ, for suitably large N it is $R_{\bar{\zeta}(\delta)-\gamma,N}(\rho) \geqslant R_{\zeta(N,\delta),N}(\rho)$, and consequently, for suitably large N it is $R_{\bar{\zeta}(\delta)/2,N}(\rho) \geqslant R_{\zeta(N,\delta),N}(\rho)$. In turn, this implies

$$R_{\bar{\zeta}(\delta)/2}(\rho) \geqslant \limsup_{N\to\infty} R_{\zeta(N,\delta),N}(\rho) \geqslant R^C_\delta(\rho),$$

and finally, being $\bar{\zeta}(\delta)$ increasing as a function of δ, taking the limit for $\delta\to 0$ one has some value $\varepsilon \geqslant 0$ such that

$$I(\rho) \geqslant \lim_{\zeta\to\varepsilon} R_\zeta(\rho) = \lim_{\delta\to 0} R_{\bar{\zeta}(\delta)}(\rho) \geqslant I^C(\rho).$$

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 $0 < \delta \leqslant \delta_0$. This means that for every $\gamma > 0$ and every N_0 there exists $N \geqslant N_0$ such that $\zeta(N,\delta) < \gamma$ for all $0 < \delta \leqslant \delta_0$. 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

$$\sup_{\mathrm{C},\{\psi_i\}} \sum_i \|[(\mathscr{D}\mathscr{E}-\mathscr{I})\boxtimes\mathscr{I}_{\mathrm{C}}]\psi_i\|_{\mathrm{op}} < \gamma,$$
$$\sup_{\mathrm{C},\{\psi_i\},\{a_j\}} L^{-1}|I(\mathrm{X}:\mathrm{Y}) - I(\mathrm{X}:\tilde{\mathrm{Y}})| > \delta,$$

where L has been introduced in Definition V.1. First, 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 $I(\mathrm{X}:\mathrm{Y}) = I(\mathrm{X}:\tilde{\mathrm{Y}}) = 0$, since $I(\mathrm{A}:\mathrm{B}) \leqslant \min\{H(\mathrm{A}), H(\mathrm{B})\}$. The minimum relevant value of L is thus $\log_2 3$. Now according to Theorem 2 in [28], for $\|\mathbf{p}-\mathbf{q}\| < \gamma < 1 - 1/mn$ one has

$$\begin{aligned} L^{-1}|I(\mathrm{X}:\mathrm{Y}) - I(\mathrm{X}':\mathrm{Y}')| &\leqslant 3\gamma + 3L^{-1}H_2(\gamma)\\ &\leqslant 3\gamma + \frac{3}{\log_2 3}H_2(\gamma),\end{aligned}$$

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 C: PROOF OF LEMMA V.2

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 optimization problem has a compact domain. Let now $\{\Psi_{\mathbf{i}}\}$ be the decomposition of $\rho^{\boxtimes N}$ defined by

$$\Psi_{\mathbf{i}} := \Psi_{i_1}\boxtimes\Psi_{i_2}\boxtimes\cdots\boxtimes\Psi_{i_N},$$

and Ψ_i be the decomposition that maximizes $I(\mathrm{X}_0:\mathrm{Y}_0)$, with m_0 outcomes. Let now $\{b_j\}$ be the observation test on N copies of the system that maximizes $I(\mathrm{X}:\mathrm{Y})$ 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 $I(\mathrm{X}:\tilde{\mathrm{Y}})$ no larger than $I(\mathrm{X}:\mathrm{Y})$. Thus we can write

$$\begin{aligned}\delta &> \frac{I(\mathrm{X}:\mathrm{Y}) - I(\mathrm{X}:\tilde{\mathrm{Y}})}{\log_2 m_0^N D(N) - 1}\\ &\geqslant \frac{I(\mathrm{X}:\mathrm{Y}) - I(\mathrm{X}:\tilde{\mathrm{Y}})}{N\log_2 m_0 D_0 + \log_2 k},\end{aligned}$$

where in the first bound we used the fact that the number of outcomes for the observation test maximizing 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 $D(N) \leqslant kD_0^N$. Now, by definition of $I(\mathrm{X}:\mathrm{Y})$ we have $I(\mathrm{X}:\mathrm{Y}) \geqslant NI(\mathrm{X}_0:\mathrm{Y}_0)$, while by the result of Theorem 2 in [29] we have

$$I(X:\tilde{Y}) \leqslant \log_2 D(M) \leqslant \log_2 k' + M\log_2 D_1,$$

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:

$$\delta > \frac{NI(X_0:Y_0) - M\log_2 D_1 - \log_2 k'}{N\log_2 m_0 D_0 + \log_2 k},$$

and consequently

$$\frac{M}{N}\frac{\log_2 D_1+\log_2 k'/M}{\log_2 m_0D_0+\log_2 k/N}+\delta>\frac{I(X_0:Y_0)}{\log_2 m_0D_0+\log_2 k/N}.$$

In particular, if the scheme $(\mathscr{E},\mathscr{D})$ has the minimum M for fixed N, δ we can then conclude that

$$R^C_{\delta,N}\frac{\log_2 D_1}{\log_2 m_0D_0+\frac{\log_2 k}{N}}+\delta+\frac{1}{N}\frac{\log_2 k'}{\log_2 m_0D_0+\frac{\log_2 k}{N}}>\frac{I(X_0:Y_0)}{\log_2 m_0D_0+\frac{\log_2 k}{N}}.$$

Taking the limit superior for $N\to\infty$ on both sides we have

$$R^C_\delta\frac{\log_2 D_1}{\log_2 m_0D_0}+\delta\geqslant\frac{I(X_0:Y_0)}{\log_2 m_0D_0},$$

and finally, in the limit $\delta\to 0$ we obtain

$$I^C(\rho)\frac{\log_2 D_1}{\log_2 m_0D_0}\geqslant\frac{I(X_0:Y_0)}{\log_2 m_0D_0},$$

namely,

$$I^C(\rho)\geqslant\frac{I(X_0:Y_0)}{\log_2 D_1}.$$

For a mixed state, $I(X_0:Y_0)>0$, and this implies the thesis.

[1] C. E. Shannon, Communication in the presence of noise, Proc. IRE **37**, 10 (1949).

[2] B. Schumacher, Quantum coding, Phys. Rev. A **51**, 2738 (1995).

[3] M. Koashi and N. Imoto, Compressibility of Quantum Mixed-State Signals, Phys. Rev. Lett. **87**, 017902 (2001).

[4] H. Barnum, C. M. Caves, C. A. Fuchs, R. Jozsa, and B. Schumacher, On quantum coding for ensembles of mixed states, J. Phys. A: Math. Gen. **34**, 6767 (2001).

[5] Z. B. Khanian and A. Winter, Entanglement-assisted quantum data compression, in *2019 IEEE International Symposium on Information Theory (ISIT)* (IEEE, Piscataway, NJ, 2019), pp. 1147–1151.

[6] B. Schumacher and M. Westmoreland, *Quantum Processes Systems, and Information* (Cambridge University Press, Cambridge, 2010).

[7] L. Hardy and W. K. Wootters, Limited holism and real-vector-space quantum theory, Found. Phys. **42**, 454 (2012).

[8] G. M. D'Ariano, F. Manessi, P. Perinotti, and A. Tosini, Fermionic computation is non-local tomographic and violates monogamy of entanglement, EPL (Europhys. Lett.) **107**, 20009 (2014).

[9] G. M. D'Ariano, F. Manessi, P. Perinotti, and A. Tosini, The Feynman problem and fermionic entanglement: Fermionic theory versus qubit theory, Int. J. Mod. Phys. A **29**, 1430025 (2014).

[10] G. M. D'Ariano, G. Chiribella, and P. Perinotti, *Quantum Theory from First Principles: An Informational Approach* (Cambridge University Press, Cambridge, 2017).

[11] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Probabilistic theories with purification, Phys. Rev. A **81**, 062348 (2010).

[12] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Informational derivation of quantum theory, Phys. Rev. A **84**, 012311 (2011).

[13] G. M. D'Ariano, M. Erba, and P. Perinotti, Classical theories with entanglement, Phys. Rev. A **101**, 042118 (2020).

[14] L. Hardy, Disentangling nonlocality and teleportation, arXiv:quant-ph/9906123 (1999).

[15] R. W. Spekkens, Evidence for the epistemic view of quantum states: A toy theory, Phys. Rev. A **75**, 032110 (2007).

[16] J. Barrett, Information processing in generalized probabilistic theories, Phys. Rev. A **75**, 032304 (2007).

[17] H. Barnum, J. Barrett, L. Orloff Clark, M. Leifer, R. Spekkens, N. Stepanik, A. Wilce, and R. Wilke, Entropy and information causality in general probabilistic theories, New J. Phys. **12**, 033024 (2010).

[18] A. J. Short and S. Wehner, Entropy in general physical theories, New J. Phys. **12**, 3023 (2010).

[19] G. Kimura, K. Nuida, and H. Imai, Distinguishability measures and entropies for general probabilistic theories, Rep. Math. Phys. **66**, 175 (2010).

[20] G. M. D'Ariano, P. Perinotti, and A. Tosini, Information and disturbance in operational probabilistic theories, Quantum **4**, 363 (2020).

[21] S. Popescu and D. Rohrlich, Quantum nonlocality as an axiom, Found. Phys. **24**, 379 (1994).

[22] G. M. D'Ariano and A. Tosini, Testing axioms for quantum theory on probabilistic toy-theories, Quant. Info. Proc. **9**, 95 (2010).

[23] P. Perinotti, A. Tosini, and L. Vaglini, Shannon theory for quantum systems and beyond: Information compression for fermions, arXiv:2106.04964 (2021).

[24] D. Gross, M. Müller, R. Colbeck, and O. C. O. Dahlsten, All Reversible Dynamics in Maximally Nonlocal Theories are Trivial, Phys. Rev. Lett. **104**, 080402 (2010).

[25] P. Perinotti, Cellular automata in operational probabilistic theories, Quantum **4**, 294 (2020).

[26] C. Fuchs and J. van de Graaf, Cryptographic distinguishability measures for quantum-mechanical states, IEEE Trans. Inf. Theory **45**, 1216 (1999).

[27] G. M. D'Ariano, M. Erba, and P. Perinotti, Classicality without local discriminability: Decoupling entanglement and complementarity, Phys. Rev. A **102**, 052216 (2020).

[28] Z. Zhang, Estimating mutual information via Kolmogorov distance, IEEE Trans. Inf. Theory **53**, 3280 (2007).

[29] S. Fiorini, S. Massar, M. K. Patra, and H. R. Tiwary, Generalized probabilistic theories and conic extensions of polytopes, J. Phys. A: Math. Theor. **48**, 025302 (2015).