Entropy in general physical theories

New Journal of Physics
The open-access journal for physics

Entropy in general physical theories

Anthony J Short[1,3] **and Stephanie Wehner**[2,3]
[1] DAMTP, Centre for Mathematical Sciences, Wilberforce Road,
Cambridge CB3 0WA, UK
[2] Institute for Quantum Information, Caltech, Pasadena, CA 91125, USA
E-mail: ajs256@cam.ac.uk and wehner@caltech.edu

New Journal of Physics **12** (2010) 033023 (34pp)
Received 15 October 2009
Published 15 March 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/3/033023

Abstract. Information plays an important role in our understanding of the physical world. Hence we propose an entropic measure of information for *any* physical theory that admits systems, states and measurements. In the quantum and classical worlds, our measure reduces to the von Neumann and Shannon entropies, respectively. It can even be used in a quantum or classical setting where we are only allowed to perform a limited set of operations. In a world that admits superstrong correlations in the form of non-local boxes, our measure can be used to analyze protocols such as superstrong random access encodings and the violation of 'information causality'. However, we also show that in such a world *no* entropic measure can exhibit all the properties we commonly accept in a quantum setting. For example, there exists *no* 'reasonable' measure of conditional entropy that is subadditive. Finally, we prove a coding theorem for some theories that is analogous to the quantum and classical settings, providing us with an appealing operational interpretation.

[3] Authors to whom any correspondence should be addressed.

Contents

1. Introduction

The understanding of information in classical and quantum physics has helped scientists to shed light on the fundamental nature of these theories. Indeed, it has even been suggested that quantum theory could be more naturally formulated in terms of its information-theoretic properties [5, 7, 10, 16]. Yet, we have barely scratched the surface of understanding the role of information in the natural world. To gain a deeper understanding of the role of information in physical systems and to help explain *why* nature is quantum, it is sometimes instructive to take a step back and view quantum mechanics in a much broader context of possible physical theories. Many examples are known that indicate that if our world were only slightly different, our ability to perform information processing tasks could change dramatically [2, 6, 15, 26, 33, 35, 37, 39].

However, before we can hope to really investigate general theories from the perspective of information processing, we need to first find a way to quantify information. In the quantum and the classical worlds, this can be done using the von Neumann and the Shannon entropies, respectively, which capture our notions of information and uncertainty in an intuitive way. These quantities have countless practical applications and have played an important role in understanding the power of such theories with respect to information processing.

Here, we propose a measure of information that applies to *any* physical theory[4] that admits the minimal notions of finite physical *systems*, their *states* and the probabilistic outcomes of *measurements* performed on them. Many such theories have been suggested, each of which shares some aspects with quantum theory and yet has important differences. For example, we might consider quantum mechanics itself with a limited set of allowed measurements, quantum mechanics in a real Hilbert space, generalized probabilistic theories [1, 3], general C*-algebraic theories [10], box world [32] (a theory admitting all non-signalling correlations [27, 42], previously called generalized non-signalling theory [3]), classical theories with an epistemic restriction [34] or theories derived by relaxing uncertainty relations [35].

1.1. A measure of information

1.1.1. Entropy. We propose an entropic measure of information $\widehat{\mathrm{H}}$ that can be used in any such theory in section 4.1. We will show that our measure reduces to the von Neumann and the Shannon entropies in the quantum and classical settings, respectively. In addition, we show that it shares many of their appealing intuitive properties. For example, we show that the quantity is always positive and bounded for the finite systems we consider. This provides us with a notion that each system has some maximum amount of information that it can contain. Furthermore, we might expect that mixing increases entropy, i.e. that the entropy of a probabilistic mixture of states cannot be less than the average entropy of its components. This is indeed the case for our entropic quantity. Another property that is desirable from a useful measure of information is that it should take on a similar value for states that are ‘close’, in the sense that there exists no way to tell them apart very well. This is the case for the von Neumann and the Shannon entropies, and also for our general entropic quantity, given one extra minor assumption. Finally, when considering two different systems A and B, one may consider how the entropy of the joint system AB relates to the entropy of the individual systems. It is intuitive that our uncertainty about the entire system AB should not exceed the sum of our uncertainties about A and B

[4] Although for simplicity, we will restrict our analysis to objects that are finite in an appropriate sense.

individually. This property is known as subadditivity and is obeyed by our measure of entropy given one additional reasonable assumption on the physical theory. Our entropic quantity thus behaves in very intuitive ways. Yet, we will see that there exist physical theories for which it is not strongly subadditive, unlike in quantum mechanics.

Of course, there are multiple ways to quantify information and we discuss our choice by examining some alternatives and possible extensions such as notions of accessible information, relative entropy as well as Rényi entropic quantities in sections 4.3 and 4.4.

1.1.2. Conditional entropy and mutual information. Clearly, it is also desirable to capture our uncertainty about some system A *conditioned* on the fact that we have access to another system B. This is captured by the *conditional* entropy, for which we provide two definitions in section 4.2, which are both interesting and useful in their own right. Based on such definitions, we also define notions of mutual information that allow us to quantify the amount of information that two systems hold about each other. Our first definition of conditional entropy is analogous to the quantum setting, and indeed reduces to the conditional von Neumann entropy in a quantum world. This is an appealing feature and opens the possibility of interesting operational interpretations of this quantity as in a quantum setting [20, 21]. Yet, we will see that there exists a theory (called box world) for which not only the subadditivity of the conditional entropy is violated but also conditioning *increases* entropy. Intuitively, we would not expect to grow more uncertain when given additional information, which we could always choose to ignore.

We will hence also introduce a second definition of conditional entropy, which does not reduce to the von Neumann entropy in the quantum world. However, it has the advantage that in *any* theory, conditioning *reduces* our uncertainty, as we would intuitively expect when taking an operational viewpoint. Nevertheless, even our second definition of the conditional entropy violates subadditivity.

1.1.3. Possible properties of the conditional entropy. Naturally, one might ask whether the fact that both our definitions of the conditional entropy violate subadditivity is simply a shortcoming of our definitions. In section 6, we therefore examine what properties any ‘reasonable’ measure of conditional entropy can have in principle. By ‘reasonable’ here we mean that if, given access to a system B, we have no uncertainty about some classical information A, then the quantity is ‘0’, otherwise it is positive (or even nonzero). We show that under this simple assumption there exists *no* measure of conditional entropy in box world that is subadditive or obeys a chain rule.

1.2. Examples

To give some idea of how our entropies can be used outside of quantum theory, we examine a very simple example in box world in section 5, which illustrates all the peculiar properties our entropies can have. This is based on a task in which Alice must produce an encoding of a string x, such that Bob can retrieve any bit of his choosing with some probability [38] (known as a random access encoding). It is known that superstrong random access codes exist in box world [35], leading to a violation of the quantum bound for such encodings [23].

A similar game was used in [26] to argue that one of the defining characteristics that set the quantum world apart from other possibilities (and particularly box world) is that communication of m classical bits causes information gain of at most m bits, a principle called ‘information causality'. In section 7, we examine this statement using our entropic quantity. We note that it is the failure of subadditivity of conditional entropy in box world that leads to a violation of the inequality quantifying 'information causality' given in [26]. We conclude our examples by discussing the definition of 'information causality' more generally.

1.3. A coding theorem

In the classical as well as the quantum setting, the Shannon and the von Neumann entropies have appealing operational interpretations as they capture our ability to compress information. In section 8, we show that the quantity $\widehat{\mathrm{H}}(\cdot)$ has a similar interpretation for some physical theories. When defining entropy we have chosen to restrict ourselves to a minimal set of assumptions, only assuming that a theory would have some notion of states and measurements. To consider compressing a state or indeed decoding it again, however, we need to know a little more about our theory. In particular, we first have to define a notion of 'size' for any compression procedure to make sense. Secondly, we need to consider what kind of encoding and decoding operations we allowed to perform. Given these ideas and several additional assumptions on our physical theory, we prove a simple coding theorem.

1.4. Outline

In section 2, we introduce a framework for describing states, measurements and transformations in general physical theories, followed by some examples in section 3. In section 4, we then define our entropic measures of information that can be applied in any theory. Examples of how these entropies can be applied in box world can be found in section 5. In section 6, we examine what properties we can hope to expect from a conditional entropy in box world. Section 7 investigates the notion of 'information causality' in our framework and finally we show a coding theorem for many theories in section 8. We conclude with many open questions in section 9.

2. An operational framework for physical theories

We now present a simple framework, based on minimal operational notions (such as systems, states, measurements and probabilities), that encompasses both classical and quantum physics, as well as more novel possibilities (such as 'box world') [1, 3, 11, 16]. Our approach is similar to that in [1]; however, it is slightly more general as it does not assume that all measurements that are mathematically well defined are physically implementable or that joint systems can be characterized by local measurements.

2.1. Single systems and states

Firstly, we will assume that there is a notion of discrete physical *systems*. With each system A we associate a set of allowed *states* $\mathcal{S}_{\mathrm{A}}$, which may differ for each system. Furthermore, we assume that we can prepare arbitrary mixtures of states (for example, by tossing a biased coin, and preparing a state dependent on the outcome) and therefore take $\mathcal{S}_{\mathrm{A}}$ to be a convex set, with $s_{\mathrm{mix}} = ps_1 + (1-p)s_2$ denoting the state that is a mixture of s_1 with probability p and s_2 with probability $1-p$. To characterize when two states are the same, or close to each other, we first need to introduce the notion of measurements.

2.2. Measurements

Secondly, we thus assume that on each system A, we can perform a certain set of allowed measurements $\mathcal{E}_{\mathrm{A}} = \{\mathbf{e}\}$. If the system A is clear from the context, we will omit the subscripts and simply write $\mathcal{E}$ and $\mathcal{S}$.

With each measurement $\mathbf{e}$ we associate a set of outcomes $\mathcal{R}_{\mathbf{e}}$, which for simplicity of exposition we take to be finite. When a particular measurement is performed on a system, the probability of each outcome should be determined by its state. We therefore associate each possible outcome $r \in \mathcal{R}_{\mathbf{e}}$ with a functional $e_r : \mathcal{S} \to [0, 1]$, such that $e_r(S)$ is the probability of obtaining outcome r given state S. We refer to such a functional as an *effect*. To ensure that measurement behaves according to our intuition when applied to mixed states, we require that $e_r(S_{\mathrm{mix}}) = p\, e_r(S_1) + (1 - p)e_r(S_2)$. This means that each effect can be taken to be *linear*[5]. In order for the probabilities of all measurement outcomes to sum to one, we also require that

$$\sum_{r \in \mathcal{R}_{\mathbf{e}}} e_r = u, \tag{1}$$

where u is the *unit effect*, which has the property that $u(S) = 1$ for all $S \in \mathcal{S}$. We can thus characterize a measurement $\mathbf{e}$ as a set of outcome/effect pairs[6]

$$\mathbf{e} = \left\{ (r, e_r) \mid r \in \mathcal{R}_{\mathbf{e}} \quad \text{and} \quad \sum_r e_r = u \right\}. \tag{2}$$

We write $\mathbf{e}(S)$ for the probability distribution over outcomes when $\mathbf{e}$ is performed on a state S. Note that in this general framework, not all measurements that are mathematically well defined need be part of a particular physical theory.

One measurement can be equivalent to, or strictly more informative than, another. Consider two measurements $\mathbf{e}$ (with outcomes $\mathcal{R}_{\mathrm{e}}$ and effects e_r) and $\mathbf{f}$ (with outcomes $\mathcal{R}_{\mathrm{f}}$ and effects f_r), for which there exists a map $M : \mathcal{R}_{\mathrm{e}} \to \mathcal{R}_{\mathrm{f}}$ such that

$$\sum_{\{r\,:\,M(r)=r'\}} e_r = f_{r'}, \qquad \forall\, r' \in \mathcal{R}_{\mathrm{f}}. \tag{3}$$

If M is one-to-one it corresponds to a *re-labelling* of the outcomes. Otherwise, we say that $\mathbf{f}$ is a *coarse-graining* of $\mathbf{e}$ (or alternatively that $\mathbf{e}$ is a *refinement* of $\mathbf{f}$). Because we can always re-label the outcomes of an experiment according to any map M, we assume that $\mathcal{E}$ is closed under re-labelling and coarse-graining. This implies that $\mathcal{E}$ always contains the trivial measurement $\mathbf{u}$ (with one outcome corresponding to effect u).

A refinement/coarse-graining is trivial if

$$e_r \propto f_{M(r)}, \qquad \forall\, r \in \mathcal{R}_{\mathrm{e}}. \tag{4}$$

In this case, the measurement of $\mathbf{e}$ is equivalent to performing $\mathbf{f}$ and obtaining r', then outputting a randomly selected r satisfying $M(r) = r'$ (where the distribution depends on the

[5] Strictly speaking, we only need the functional to act linearly on mixtures, requiring it to be affine. However, it is helpful to extend it to full linearity to deal with unnormalized states.

[6] Note that here we do not describe a measurement as a set of effects, as this rules out the possibility of two or more measurement outcomes corresponding to the same effect and makes it harder to discuss coarse-graining, re-labelling or expectation values.

proportionality constant in (4)). Hence the two measurements are equally informative about the state. In contrast, when $\mathbf{e}$ is a nontrivial refinement of $\mathbf{f}$, it offers strictly more information about the state, and in this case we write $\mathbf{e} \succ \mathbf{f}$. A subset of measurements of particular importance are the *fine-grained* measurements $\mathcal{E}^* \subseteq \mathcal{E}$, which have no nontrivial refinements and are therefore optimal for gathering information about the state. Formally,

$$\mathbf{e} \in \mathcal{E}^* \;\Leftrightarrow\; \nexists\, \mathbf{f} \in \mathcal{E} \;: \mathbf{f} \succ \mathbf{e}. \tag{5}$$

We will also call an effect $\mathbf{e}$ fine-grained if it is part of a fine-grained measurement. We assume that $\mathcal{E}^*$ is non-empty (i.e. that there exists at least one finite outcome fine-grained measurement). In quantum and classical theory, this restricts us to the finite-dimensional case.

2.3. Transformations

As well as preparing states and performing measurements, it may be possible to perform transformations on a system. As in the case of effects, in order to behave reasonably when applied to mixed states, a transformation must correspond to a linear map $T : \mathcal{S}_{\mathrm{A}} \to \mathcal{S}_{\mathrm{A}'}$ taking allowed states to allowed states (although the input and output systems may be of a different type). For each type of system, there will be some set of allowed transformations $\mathcal{T}$.

We assume that the identity transformation I is allowed and that the composition of two allowed transformations is allowed (as long as the system output by the first transformation is of the same type as the input to the second). Furthermore, it must be the case that any allowed transformation followed by an allowed measurement is an allowed measurement.

We can also combine the notion of transformation with that of measurement in a natural way to represent non-destructive measurements [3, 11]. To incorporate non-destructive measurements, define the sub-normalized states $\tilde{\mathcal{S}} = \{pS | 0 \leqslant p \leqslant 1, S \in \mathcal{S}\}$. A measurement can then be described by assigning a sub-normalized transformation $t_r : \mathcal{S} \to \tilde{\mathcal{S}}'$ to each outcome r. Result r occurs with probability $p_r = u(t_r(s))$ and the post-measurement state is $s_r = t_r(s)/p_r$. However, we will not need such constructions in the main part of this paper.

2.4. Relations between states

Having introduced measurements, we can now define what it means for two states to be equal. Given that we are taking an operational viewpoint, we adopt the intuitive notion that two states $S_1, S_2 \in \mathcal{S}$ are equal if and only if there exists no measurement that distinguishes them. That is,

$$\forall S_1, S_2 \in \mathcal{S}, \quad S_1 = S_2 \;\Leftrightarrow\; \forall\, \mathbf{e} \in \mathcal{E} \;:\; \mathbf{e}(S_1) = \mathbf{e}(S_2). \tag{6}$$

We can also define a natural measure of distance for states $S_0, S_1 \in \mathcal{S}$ that directly relates to the probability that we can distinguish these states using measurements available in our theory, in analogy to the quantum setting [22]. Suppose that we are given either S_0 or S_1 with equal probability and perform a measurement $\mathbf{e}$ to distinguish the two cases. Note that the above implies that any theory that admits at least two possible states has at least one measurement $\mathbf{e}$ with two possible outcomes. Furthermore, any such theory must have a measurement $\mathbf{e}$ with exactly two outcomes since any theory admits arbitrary coarse-grainings of measurements. We will base our decision on the maximum likelihood rule; that is, when we obtain outcome r, we will conclude that we received state S_0 if $e_r(S_0) > e_r(S_1)$ and S_1 otherwise. The probability of distinguishing the two states using measurement **e** is then given by

$$p^{\mathbf{e}}_{\text{succ}} = \frac{1}{2} + \frac{\mathcal{C}(\mathbf{e}(S_0), \mathbf{e}(S_1))}{2}, \tag{7}$$

where $\mathcal{C}(\mathbf{e}(S_0), \mathbf{e}(S_1)) = \frac{1}{2}\sum_{r\in\mathcal{R}_{\mathrm{e}}} |e_r(S_0) - e_r(S_1)|$ is the classical statistical distance between the probability distributions $\mathbf{e}(S_0)$ and $\mathbf{e}(S_1)$. We now define the distance as

$$\mathcal{D}(S_0, S_1) := \sup_{\mathbf{e}} \mathcal{C}(\mathbf{e}(S_0), \mathbf{e}(S_1)). \tag{8}$$

By the above, we see that this measure of distance has an appealing operational interpretation because it directly captures our ability to distinguish the two states S_0 and S_1 using any available measurement (see appendix A, lemma A.1 for details). In the quantum setting, it thus directly reduces to the well-known trace distance.

2.5. *Multi-partite systems*

Suppose that we have two systems A and B, each of which may admit different sets of states and measurements. We allow that two individual systems can be combined into a *composite* system AB, which we can treat as a new type of system having its own set of allowed states, measurements, and transformations just as in the single-system case. However, these sets must bear some relation to those of the component subsystems.

With respect to states, we would like it to be possible to independently prepare any state $S_{\mathrm{A}} \in \mathcal{S}_{\mathrm{A}}$ of system A and $S_{\mathrm{B}} \in \mathcal{S}_{\mathrm{B}}$ of system B. This corresponds to a *product state* of the composite system, which we denote by $S_{\mathrm{AB}} = S_{\mathrm{A}} \otimes S_{\mathrm{B}} \in \mathcal{S}_{\mathrm{AB}}$. Note that at this point we have not proved that $\otimes$ corresponds to a tensor product in the usual sense[7], but we would nevertheless expect that it is distributive for mixtures and associative. We make use of the standard terminology that states are *separable* if they can be written as a mixture of product states, and *entangled* otherwise. To avoid excessive subscripts when dealing with multiple systems, we will usually refer to the state of systems AB and B directly by these letters, rather than the more cumbersome S_{AB} and S_{A} (e.g. $\mathbf{e}(S_{\mathrm{AB}}) = \mathbf{e}(\mathrm{AB})$, etc).

Similarly, we would expect to be able to perform a measurement $\mathbf{e} \in \mathcal{E}_{\mathrm{A}}$ and $\mathbf{f} \in \mathcal{E}_{\mathrm{B}}$, giving a product measurement that we denote by $\mathbf{g} = \mathbf{e} \otimes \mathbf{f} \in \mathcal{E}_{\mathrm{AB}}$ (with outcome set $\mathcal{R}_{\mathbf{g}} = \mathcal{R}_{\mathbf{e}} \times \mathcal{R}_{\mathbf{f}}$ and effects $g_{ij} = e_i \otimes f_j$). By considering coarse-graining and tri–partite systems, we would again expect $\otimes$ to be distributive and associative. When applying a product measurement to a product state, we furthermore require that

$$(e_i \otimes f_j)(\mathrm{A} \otimes \mathrm{B}) = e_i(\mathrm{A}) f_j(\mathrm{B}). \tag{9}$$

When considering multiple systems, we can consider what happens if we measure only some of these systems. Note that this means that we perform a measurement consisting of a unit effect on some of these systems. This only makes sense if marginal states are well defined and hence we assume that even when a bipartite state is entangled, each part is an allowed marginal state. We can thus have

$$\forall\, (\mathrm{AB}) \in \mathcal{S}_{\mathrm{AB}}, \quad \exists \mathrm{A} \in \mathcal{S}_{\mathrm{A}} : \forall\, \mathbf{e} \in \mathcal{E}_{\mathrm{A}}, \quad \mathbf{e}(\mathrm{A}) = (\mathbf{e} \otimes \mathbf{u})(\mathrm{AB}). \tag{10}$$

[7] It is possible to prove this with the additional assumption that multi-partite systems are completely characterized by product measurements [1, 3]. However, this rules out some potentially reasonable theories, such as quantum theory in a real Hilbert space, and we will not need to make this additional assumption here.

Furthermore, in the case where B performs a measurement on his subsystem and obtains result r (corresponding to an effect e_r), we would expect A's subsystem to 'collapse' to an allowed state $\mathrm{A}_{|r} \in \mathcal{S}_\mathrm{A}$. We will denote such a state as

$$\mathrm{A}_{|r} = \frac{(I \otimes e_r)(\mathrm{AB})}{e_r(\mathrm{B})}. \tag{11}$$

Finally, a crucial constraint on multi-partite systems is the existence of product transformations $T_\mathrm{A} \otimes T_\mathrm{B} \in \mathcal{T}_\mathrm{AB}$. In a variant of quantum theory in which all positive (rather than completely positive) trace-preserving maps are allowed transformations, this would prevent the existence of entangled states.

3. Example theories

In this section, we show how quantum theory and classical probability theory fit into the framework defined above and also describe the theory known as 'box world' [3, 32], which admits all non-signalling correlations [27, 42], and was one of the main motivations for this work.

3.1. Classical probability theory

In classical probability theory, a state S corresponds to a probability distribution p_i over a finite set of elements. The effects correspond to linear functionals of the form

$$e_r(S) = \sum_i q_r^i p_i \tag{12}$$

for any $q_r^i \in [0, 1]$. Note that the unit effect corresponds to $q^i = 1\,\forall\, i$. Normalization of measurements, therefore, requires $\sum_r q_r^i = 1\,\forall\, i$. Transformations correspond to stochastic maps.

3.2. Quantum theory

In quantum theory, the convex set of states are the density operators $S = \rho$ (trace-1 positive operators), and effects correspond to linear functionals of the form

$$e_r(S) = \mathrm{tr}(\rho E_r), \tag{13}$$

where E_r is a positive operator. All measurements satisfying the normalization constraint

$$\sum_r e_r = u \Rightarrow \sum_r E_r = I \tag{14}$$

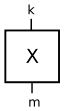
are allowed, and the fine-grained measurements are those for which all E_r are rank 1 operators. The allowed transformations represent completely positive trace-preserving maps [24].

3.3. Restricted quantum/classical theories

Note that unlike other approaches [1, 3], our framework also encompasses real Hilbert space quantum mechanics. Furthermore, because we do not assume that all well-defined operations are physically realizable, it can be used to study quantum or classical theory with a restricted set of states, measurements and transformations (for an interesting example in the classical case, consider Spekkens' toy model [34]). The entropies we would assign in such cases would differ from the standard von Neumann entropy and may be interesting to study.

3.4. Box world

In box world, the state of a single system X corresponds to a conditional probability distribution $S = P(x_{\text{out}}|x_{\text{in}})$, where x_{in} and x_{out} are the elements of a finite set of 'inputs' and 'outputs', respectively. The idea is that there is a special set of measurements on each system represented by x_{in} (referred to as *fiducial* measurements), and that any probability distribution for these measurements corresponds to an allowed state. We represent a system X with k possible inputs x_{in} and m possible outputs x_{out} by



In the special case where there is only one possible input, the conditional probability distribution reduces to the standard unconditional probability distribution $P(x_{\text{out}})$, and we omit the input line to the box in the diagram. Thus box world contains classical probability theory as a special case, and we will use such *classical boxes* to represent classical information in our treatment of information-theoretic protocols in box world.

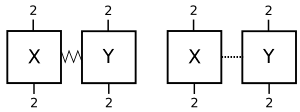
A multi-partite state in box world corresponds to a joint conditional probability distribution $P(x_{\text{out}}^1 x_{\text{out}}^2 \ldots x_{\text{out}}^N | x_{\text{in}}^1 x_{\text{in}}^2 \ldots x_{\text{in}}^N)$ with a separate input and output for each system. Aside from the usual constraints of normalization and positivity, the allowed states must also satisfy the non-signalling conditions: that the marginal probability distribution obtained by summing over x_{out}^k,

$$\sum_{x_{\text{out}}^k} P(x_{\text{out}}^1 \ldots x_{\text{out}}^k \ldots x_{\text{out}}^N | x_{\text{in}}^1 \ldots x_{\text{in}}^k \ldots x_{\text{in}}^N), \tag{15}$$

is independent of x_{in}^k for all k. This means that the other parties cannot learn anything about a distant party's measurement choice from their own measurement results. A bipartite state of particular interest is the PR-box state [27]–[29], for which all inputs and outputs are binary, and the probability distribution is

$$P_{\text{PR}}(x_{\text{out}}^1 x_{\text{out}}^2 | x_{\text{in}}^1 x_{\text{in}}^2) = \begin{cases} \frac{1}{2} & : \quad x_{\text{out}}^1 \oplus x_{\text{out}}^2 = x_{\text{in}}^1 \cdot x_{\text{in}}^2, \\ 0 & : \quad \text{otherwise}, \end{cases} \tag{16}$$

where $\oplus$ denotes addition modulo 2. This state is 'more entangled' than any quantum state, yielding correlations that achieve the maximum possible value of 4 for the Clauser–Horne–Shimony–Holt (CHSH) expression [9], compared to $\leqslant 2\sqrt{2}$ for quantum theory (Tsirelson's bound [36]), and $\leqslant 2$ for classical probability theory. We represent entanglement between systems in box world by a zigzag line between them, and classical correlations (i.e. separable but non-product states) by a dotted line.



In box world, we allow all mathematically well-defined measurements and transformations to be physically implemented. Writing $\vec{x}_{\text{out}} = (x^1_{\text{out}}, x^2_{\text{out}}, \ldots, x^N_{\text{out}})$ and $\vec{x}_{\text{in}} = (x^1_{\text{in}}, x^2_{\text{in}}, \ldots, x^N_{\text{in}})$, all effects take the form

$$e_r(S) = \sum_{\vec{x}_{\text{out}},\vec{x}_{\text{in}}} Q_r(\vec{x}_{\text{out}}|\vec{x}_{\text{in}}) P(\vec{x}_{\text{out}}|\vec{x}_{\text{in}}), \tag{17}$$

where $Q_r(x_{\text{out}}|\vec{x}_{\text{in}})$ can be taken to be positive [3]. The effect $e^{\vec{x}'_{\text{in}}}_{\vec{x}'_{\text{out}}}$ corresponding to performing joint fiducial measurements $\vec{x}'_{\text{in}}$ and obtaining results $\vec{x}'_{\text{out}}$ is represented by $Q_{\vec{x}'_{\text{out}}}(\vec{x}_{\text{out}}|\vec{x}_{\text{in}}) = \delta_{x_{\text{in}}x'_{\text{in}}} \delta_{x_{\text{out}},x'_{\text{out}}}$. Because of the positivity of Q_r, any effect can be expressed as a weighted sum of such fiducial measurement effects. It follows that a measurement is fine-grained if and only if each of its effects is proportional to some $e^{\vec{x}_{\text{in}}}_{\vec{x}_{\text{out}}}$ and that products of fine-grained measurements are themselves fine-grained.

4. Generalized entropies

The Shannon entropy $H(\vec{p}) = -\sum_i p_i \log p_i$ and von Neumann entropy $S(\rho) = -\text{tr}(\rho \log \rho)$ are extremely useful tools for analysing information processing in a classical or quantum world. Here, we would like to define an analogous entropy for general probabilistic theories, which reduces to $H(\vec{p})$ and $S(\rho)$ for classical probability theory and quantum theory, respectively. We would also like our new entropy to retain as many of the mathematical properties of the Shannon and the von Neumann entropies as possible. Not only will this help our new entropy conform to our intuitive notions, but also it will make it easier to prove general results using these quantities, and transfer known results to the general case. Note that although we can use any base for the logarithm in the definition of the Shannon and the von Neumann entropies (as long as we are consistent), in what follows we will use base 2 (i.e. $\log = \log_2$) throughout.

4.1. Entropy

We now give a concrete definition of entropy for any physical theory, which satisfies the above desiderata. Other definitions are certainly possible, and we will consider one alternative (based on mixed state decomposition) in section 4.4. However, the following definition has many appealing properties.

Given any state $S \in \mathcal{S}$, we define its entropy $\widehat{\text{H}}(S)$ by

$$\widehat{\text{H}}(S) := \inf_{\mathbf{e}\in\mathcal{E}^*} \text{H}(\mathbf{e}(S)), \tag{18}$$

where the infimum is taken over all fine-grained measurements $\mathbf{e} \in \mathcal{E}^*$ on the state space $\mathcal{S}$ and $\text{H}(\mathbf{e}(S)) = -\sum_{r\in\mathcal{R}_{\mathbf{e}}} e_r(S) \log e_r(S)$ is the Shannon entropy of the probability distribution $\mathbf{e}(S)$

over possible outcomes of **e**. This has an intuitive operational meaning as the minimal output uncertainty of any fine-grained measurement on the system. Note that for information-gathering purposes, the best measurements are always fine-grained, and without restricting to this subset the unit measurement would always be optimal (giving zero outcome uncertainty). Furthermore, note that trivial refinements of **e** always generate a higher output entropy, so it is sufficient to only consider measurements in the infimum that have no parallel effects.

In appendix B, we prove that $\widehat{\mathrm{H}}$ retains several important properties of the Shannon and the von Neumann entropies. In particular, we show the following.

(i) *Reduction*. $\widehat{\mathrm{H}}$ reduces to the Shannon entropy for classical probability theory and the von Neumann entropy for quantum theory.

(ii) *Positivity and boundedness*. Suppose that the minimal number of outcomes for a fine-grained measurement in $\mathcal{E}^*_S$ is d. Then for all states $S \in \mathcal{S}$,

$$\log(d) \geqslant \widehat{\mathrm{H}}(S) \geqslant 0. \tag{19}$$

(iii) *Concavity*. For any $S_1, S_2 \in \mathcal{S}$ and any mixed state $S_{\mathrm{mix}} = pS_1 + (1-p)S_2 \in \mathcal{S}$:

$$\widehat{\mathrm{H}}(S_{\mathrm{mix}}) \geqslant p\widehat{\mathrm{H}}(S_1) + (1-p)\widehat{\mathrm{H}}(S_2). \tag{20}$$

(iv) *Limited subadditivity*. Consider a theory with the additional property that fine-grained measurements remain fine-grained for composite systems, i.e.

$$\mathbf{e} \in \mathcal{E}^*_{\mathrm{A}}, \quad \mathbf{f} \in \mathcal{E}^*_{\mathrm{B}} \Rightarrow \mathbf{e} \otimes \mathbf{f} \in \mathcal{E}^*_{\mathrm{AB}}. \tag{21}$$

This is true in quantum theory, classical theory and box world. When (21) holds, then for any bipartite state $\mathrm{AB} \in \mathcal{S}_{\mathrm{AB}}$ and reduced states $\mathrm{A} \in \mathcal{S}_{\mathrm{A}}$ and $\mathrm{B} \in \mathcal{S}_{\mathrm{B}}$,

$$\widehat{\mathrm{H}}(\mathrm{A}) + \widehat{\mathrm{H}}(\mathrm{B}) \geqslant \widehat{\mathrm{H}}(\mathrm{AB}). \tag{22}$$

(v) *Limited continuity*. Consider a system for which all allowed measurements have at most D outcomes or for which restricting the allowed measurements to have at most D outcomes does not change the entropy of any state. This is true in quantum theory, with $D = d = \dim(\mathcal{H})$, and also in box world and classical theory. Then we can prove an analogue of the Fannes inequality [14, 17], which says that the entropy of two states that are close does not differ too much. In particular, given $S_1, S_2 \in \mathcal{S}$ satisfying $\mathcal{D}(S_1, S_2) < 1/e$,

$$|\widehat{\mathrm{H}}(S_1) - \widehat{\mathrm{H}}(S_2)| \leqslant \mathcal{D}(S_1, S_2) \log\left(\frac{D}{\mathcal{D}(S_1, S_2)}\right). \tag{23}$$

We will also see in section 8 that $\widehat{\mathrm{H}}$ has an appealing operational interpretation as a measure of compressibility for some theories.

However, one property of the von Neumann entropy that does not carry over to $\widehat{\mathrm{H}}$ is strong subadditivity [24]. In particular, we will see in section 5 that there exists a tripartite state in box world such that

$$\widehat{\mathrm{H}}(\mathrm{ABC}) + \widehat{\mathrm{H}}(\mathrm{C}) > \widehat{\mathrm{H}}(\mathrm{AC}) + \widehat{\mathrm{H}}(\mathrm{BC}). \tag{24}$$

4.2. Conditional entropy and mutual information

4.2.1. A standard definition. Based on the entropy $\widehat{\mathrm{H}}$, we can also define a notion of conditional entropy. In analogy to the von Neumann entropy [8], we define the conditional entropy of a general bipartite state $\mathrm{AB} \in \mathcal{S}_{\mathrm{AB}}$ with reduced states $\mathrm{A} \in \mathcal{S}_{\mathrm{A}}$ and $\mathrm{B} \in \mathcal{S}_{\mathrm{B}}$ by

$$\widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B}) := \widehat{\mathrm{H}}(\mathrm{AB}) - \widehat{\mathrm{H}}(\mathrm{B}). \tag{25}$$

This has the nice property that for quantum or classical systems it reduces to the conditional von Neumann and Shannon entropies, respectively. In some theories (including quantum theory but not classical probability theory), $\widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B})$ can be negative, which is strange, but opens the way for an appealing operational interpretation as in the quantum setting [20].

However, unlike in quantum theory, we will see that $\widehat{\mathrm{H}}(\cdot|\cdot)$ has the counterintuitive property that it can *decrease* when 'forgetting' information in some probabilistic theories. In particular, the violation of strong subadditivity for $\widehat{\mathrm{H}}$ in box world implies that it is possible to obtain $\widehat{\mathrm{H}}(\mathrm{A}|\mathrm{BC}) > \widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B})$, and that $\widehat{\mathrm{H}}(\cdot|\cdot)$ is not subadditive. These properties will motivate us to consider an alternative definition of the conditional entropy below. However, we will show that no 'reasonable' entropy in box world can have all the appealing properties of the conditional von Neumann entropy.

In analogy to the quantum case, we can also define the *mutual information* via

$$\begin{aligned}\hat{\mathrm{I}}(\mathrm{A};\mathrm{B}) &:= \widehat{\mathrm{H}}(\mathrm{A}) + \widehat{\mathrm{H}}(\mathrm{B}) - \widehat{\mathrm{H}}(\mathrm{AB}) \\ &= \widehat{\mathrm{H}}(\mathrm{A}) - \widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B}) = \widehat{\mathrm{H}}(\mathrm{B}) - \widehat{\mathrm{H}}(\mathrm{B}|\mathrm{A}).\end{aligned} \tag{26}$$

This quantity will be positive whenever subadditivity holds, and reduces to the usual mutual information in the quantum and classical case. Similarly, we may define a notion of *accessible information* analogous to the quantum setting as

$$\hat{\mathrm{I}}_{\mathrm{acc}}(\mathrm{A};\mathrm{B}) := \sup_{\mathbf{e}\in\mathcal{E}_{\mathrm{A}},\mathbf{f}\in\mathcal{E}_{\mathrm{B}}} \mathrm{I}(\mathbf{e}(\mathrm{A});\mathbf{f}(\mathrm{B})), \tag{27}$$

where I is the classical mutual information.

4.2.2. An alternative definition. Given the problems observed with the previous definition in some theories, we now define a second form of conditional entropy based on $\widehat{\mathrm{H}}$, which sometimes captures our intuitive notions about information in a nicer way. For any bipartite state $\mathrm{AB} \in \mathcal{S}_{\mathrm{AB}}$ with reduced states $\mathrm{A} \in \mathcal{S}_{\mathrm{A}}$ and $\mathrm{B} \in \mathcal{S}_{\mathrm{B}}$, we define

$$\widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{B}) := \inf_{\mathbf{f}\in\mathcal{E}_{\mathrm{B}}} \sum_{j} f_j(\mathrm{B})\widehat{\mathrm{H}}(\mathrm{A}_{|j}), \tag{28}$$

where the infimum is taken over all measurements on B, and $\mathrm{A}_{|j}$ is the reduced state of the first system conditioned on obtaining measurement outcome j when performing $\mathbf{f}$ on the second system. This definition has the appealing property that conditioning on more systems always reduces the entropy, that is, $\widehat{\mathrm{H}}(\mathrm{A}) \geqslant \widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{B}) \geqslant \widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{BC})$ (see appendix C, lemma C.1), and it reduces to the conditional Shannon entropy in the classical case. Note, however, that $\widehat{\mathrm{H}}_{+}(\cdot|\cdot)$ does not reduce to the conditional von Neumann entropy in the quantum setting, as it is always positive. Furthermore, we will see in section 6 that it is not subadditive, and does not obey the usual chain rule (even though a limited form of chain rule holds in box world as we show in the appendix C.2). Nevertheless, $\widehat{\mathrm{H}}_{+}(\cdot|\cdot)$ seems quite a natural entropic quantity, and its corresponding quantum version has found an interesting application in the study of quantum correlations [12].

We can also define a corresponding information quantity via

$$\hat{\mathrm{I}}_{+}(\mathrm{A};\mathrm{B}) = \widehat{\mathrm{H}}(\mathrm{A}) - \widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{B}), \tag{29}$$

which is always positive. However, unlike $\hat{\mathrm{I}}(\mathrm{A};\mathrm{B})$, this definition is not symmetric and hence it cannot really be considered 'mutual information'. Instead, $\hat{\mathrm{I}}_{+}(\mathrm{A};\mathrm{B})$ captures the amount of information that B holds about A.

4.3. Other entropic quantities

For cryptographic purposes, such as in the setting of device-independent security for quantum key distribution, it is useful to define the following Rényi entropic variants of $\widehat{\mathrm{H}}$. More precisely, we define

$$\widehat{\mathrm{H}}_\alpha(S) := \inf_{\mathbf{e}\in\mathcal{E}^*} \mathrm{H}_\alpha(\mathbf{e}(S)), \tag{30}$$

where $\mathrm{H}_\alpha(\mathbf{e}(S)) = \frac{1}{1-\alpha}\log\left(\sum_j(\mathbf{e}(S)_j)^\alpha\right)$ is the Rényi entropy of order α. Note that $\mathrm{H}_1(S) = \mathrm{H}(S)$ (taking the limit of $\alpha \to 1$). These quantities can also be useful in order to bound the value of $\widehat{\mathrm{H}}(\cdot)$ itself as for any state $S \in \mathcal{S}$ and $\alpha < \beta$ we have $\widehat{\mathrm{H}}_\beta(S) \geqslant \widehat{\mathrm{H}}_\alpha(S)$.

To define a notion of relative entropy, we adopt a purely operational viewpoint. Suppose we are given N copies of a state S_1 or a state S_2, and let

$$S_1^N := S_1^{\otimes N},$$
$$S_2^N := S_2^{\otimes N}.$$

Classically, as well as quantumly, the relative entropy captures our ability to distinguish S_1^N from S_2^N for large N. Note that to distinguish the two cases, it is sufficient to coarse-grain any measurement to a two-outcome measurement $\mathbf{e} = \{(1, e_1), (2, e_2)\}$, where without loss of generality we associate the outcome '1' with the state S_1^N and '2' with S_2^N. Then $e_1(S_2^N)$ denotes the probability that we conclude that the state was S_2^N, when really we were given S_1^N. Similarly, $e_2(S_1^N)$ denotes the probability that we falsely conclude that the state was S_2^N. In what is called asymmetric hypothesis testing, we wish to minimize the error $e_1(S_2^N)$ while simultaneously demanding that $e_2(S_1^N)$ is bounded from above by a parameter ε. Here, we fix $\varepsilon = 1/2$. We therefore want to determine

$$p_N := \inf_{\mathbf{e}}\{e_1(S_2^N)|e_2(S_1^N) \leqslant 1/2\}. \tag{31}$$

In a quantum setting, it has been shown that the quantum relative entropy is directly related to this quantity via the quantum Stein's lemma [4, 18, 25], which states that we have

$$D(S_1||S_2) = \lim_{N\to\infty} -\frac{\log p_N}{N}. \tag{32}$$

This is a deep result giving a clear operational interpretation to the relative entropy, telling us that in the large N limit the probability of making the error p_N decreases exponentially with $D(S_1||S_2)$. Furthermore, as it is expressed in operational terms, we can simply adopt (32) as our definition of relative entropy in any theory for which the limit is well defined. Thus we recover the usual value in the quantum (and classical) case, and in all other theories we still capture the same operational interpretation.

Note also that our choice of $\varepsilon = 1/2$ was quite arbitrary, and one may consider a family of relative entropies, one for each choice of ε. In quantum theory, these are all equivalent [4], but they may yield different values in other theories.

4.4. Decomposition entropy

Although the entropy $\widehat{\mathrm{H}}$ has several appealing properties, and seems quite intuitive, it is nevertheless interesting to consider alternative notions of entropy for general theories. One seemingly natural alternative is the decomposition entropy, which measures the mixedness of a state.

There is a special subset of states $\mathcal{S}^* \subseteq \mathcal{S}$ that cannot be obtained by mixing other states:

$$S \in \mathcal{S}^* \Leftrightarrow \nexists S_1, \quad S_2 \in \mathcal{S}, \quad p \in (0, 1) : S = pS_1 + (1-p)S_2. \tag{33}$$

$\mathcal{S}^*$ form the extreme points of $\mathcal{S}$ and are referred to as *pure states* (with the remaining states being *mixed*). Suppose that any state in $\mathcal{S}$ can be decomposed into a finite sum of pure states. Then we can define the entropy of a state by the minimal Shannon entropy of its decompositions into pure states. Define a decomposition $\mathbf{D}(S)$ of a state $S \in \mathcal{S}$ as a probability distribution over the set of pure states that is nonzero for only a finite set of states $S_i \in \mathcal{S}^*$ with probabilities $p_i \in (0, 1]$ such that $\sum p_i S_i = S$. Then define the decomposition entropy as

$$\check{\mathrm{H}}(S) := \inf_{\mathbf{D}(S)} \mathrm{H}(\mathbf{D}(S)). \tag{34}$$

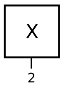
Like our previous entropy definition, we show in appendix D that $\check{\mathrm{H}}$ reduces to the Shannon and the von Neumann entropies in classical probability theory and quantum theory, respectively. However, it has a number of unappealing properties when compared with $\widehat{\mathrm{H}}$. In particular it is neither concave nor subadditive, as revealed by explicit counterexamples from box world given in appendix D.

After studying simple examples in box world, it seems that $\check{\mathrm{H}}$ is a less intuitive and helpful measure of uncertainty than $\widehat{\mathrm{H}}$. For this reason, although $\check{\mathrm{H}}$ may play an important role in discussions of entanglement or purity in many generalized theories, and may also lead to interesting operational interpretations, we do not discuss it further here.

5. Examples in box world

We now investigate how our entropic quantity $\widehat{\mathrm{H}}(\cdot)$ behaves in box world with a simple, yet illustrative, example.

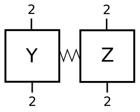
To first gain some idea of how $\widehat{\mathrm{H}}$ behaves in such a setting, consider a trivial classical system X that admits only one possible measurement and outputs two possible values $x_{\mathrm{out}} \in \{0, 1\}$ each with probability $1/2$.



Clearly, since the system admits only one possible measurement $\mathbf{e}$, we have

$$\widehat{\mathrm{H}}(X) = \mathrm{H}(\mathbf{e}(X)) = \mathrm{H}((1/2, 1/2)) = 1. \tag{35}$$

Consider now a PR-box (a bipartite system in the state (16))

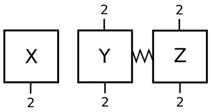


where Alice holds system Y (with binary input y_{in} and output y_{out}) and Bob holds system Z (with binary input z_{in} and output z_{out}). Note that the fine-grained measurements on the entire system correspond to a sequence of fiducial measurements on the two subsystems (where the choice of input to the second subsystem may depend on the output of the first) [3], and the outcome is the output of both measurements. The minimal entropy for the joint system can be obtained by inputting '0' into both boxes, giving outputs '00' or '11' each with probability 1/2 (in fact, any other fine-grained measurement is equally good), and the marginal states yield a random output bit for any input. Hence, we have that

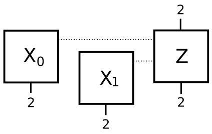
$$\widehat{\mathrm{H}}(\mathrm{Y}) = \widehat{\mathrm{H}}(\mathrm{Z}) = \widehat{\mathrm{H}}(\mathrm{YZ}) = 1. \tag{36}$$

We now consider a scenario for which it is known that PR-boxes yield an advantage over the quantum setting in terms of information processing. The basis of our example is a simple non-local game in which Alice is given a random 'parity' bit x, and has to output two bits x_0 and x_1 satisfying $x_0 \oplus x_1 = x$ (where $\oplus$ denotes addition modulo 2). Then, without receiving any communication from Alice, Bob is given a random target bit t and has to successfully output x_t [13]. This game is equivalent to the CHSH game [9, 38].

We begin with Alice having the parity bit (which we model by a classical box in the state X described above) and Alice and Bob sharing a PR-box in the state YZ.



Now Alice performs the following procedure, which corresponds to an allowed transformation in box world. She measures the parity bit X to obtain $x := x_{\text{out}}$, then uses this as the input to her part of the PR-box, setting $y_{\text{in}} = x$ and obtaining outcome y_{out}. Finally, she prepares two new classical bits $x_0 = y_{\text{out}}$ and $x_1 = x \oplus y_{\text{out}}$ (represented by classical boxes X_0, X_1). Note that because of the correlations inherent in the PR-box, the output of Bob's system will now be described by $z_{\text{out}} = y_{\text{in}} \cdot z_{\text{in}} \oplus y_{\text{out}} = (x_0 \oplus x_1) \cdot z_{\text{in}} \oplus x_0 = x_{z_{\text{in}}}$. Hence the state of $\mathrm{X}_0\mathrm{X}_1\mathrm{Z}$ after this procedure is the classically correlated state:



$$P(x_0 x_1 z_{\text{out}} | z_{\text{in}}) = \begin{cases} \frac{1}{4} & : \quad z_{\text{out}} = x_{z_{\text{in}}} \\ 0 & : \quad \text{otherwise} \end{cases} \tag{37}$$

Given any target bit t, Bob can win the game by setting $z_{\text{in}} = t$ and outputting the result $z_{\text{out}} = x_t$. We can think of Bob's system as a perfect random access encoding of the two-bit string $x_0 x_1$ [35, 38].

Consider the entropies of the state X_0X_1Z. All of the individual systems yield a random output bit, giving

$$\widehat{H}(X_0) = \widehat{H}(X_1) = \widehat{H}(Z) = 1, \tag{38}$$

and x_0 and x_1 are independent random bits, so

$$\widehat{H}(X_0X_1) = 2. \tag{39}$$

Also note that we have

$$\widehat{H}(X_0X_1Z) = 2, \tag{40}$$

since, for any input z_{in}, the output z_{out} will be perfectly correlated with one of the other bits (giving only two independent random output bits). Finally, because we can make z_{out} perfectly correlated with either of the remaining bits we have

$$\widehat{H}(X_0Z) = \widehat{H}(X_1Z) = 1, \tag{41}$$

where the optimal measurements are $z_{\text{in}} = 0$ and $z_{\text{in}} = 1$, respectively.

These entropy values all seem very intuitive (note, in contrast, that for the decomposition entropy, $\breve{H}(X_0Z) = 2$). However, they violate several natural properties of the Shannon and von Neumann entropies.

(a) Strong subadditivity. First of all, it is easy to see from the above that

$$\widehat{H}(X_0X_1Z) + \widehat{H}(Z) > \widehat{H}(X_0Z) + \widehat{H}(X_1Z), \tag{42}$$

which violates strong subadditivity. We now turn to the two possible forms of conditional entropy that we defined, where our simple example clearly illustrates their differences.

5.1. Standard conditional entropy

First of all, we consider the standard form of conditional entropy, which reduces to the von Neumann entropy in the quantum settings. By the above, we can immediately see that it has the following interesting properties.

(b) Subadditivity of the conditional entropy. Using (25) we deduce that

$$\widehat{H}(X_0|Z) = \widehat{H}(X_1|Z) = 0, \quad \widehat{H}(X_0X_1|Z) = 1, \tag{43}$$

which seems intuitive, as we can perfectly predict the output of either X_0 or X_1 (but not both) using Z. However, this yields a violation of subadditivity for the conditional entropy, as

$$\widehat{H}(X_0X_1|Z) > \widehat{H}(X_0|Z) + \widehat{H}(X_1|Z). \tag{44}$$

This may seem rather bizarre at first glance; however, we will see in section 6 that no ‘reasonable’ measure of conditional entropy in box world is sub-additive, unlike the von Neumann entropy.

It is also interesting to consider the corresponding mutual information quantities, which are

$$\hat{I}(X_0; Z) = \hat{I}(X_1; Z) = \hat{I}(X_0X_1; Z) = 1. \tag{45}$$

Again, these seem intuitive, as we can extract one bit of information about either X_0 or X_1 or the pair X_0X_1 from Z.

It may be tempting to conclude that the point at which $\widehat{H}(X_0X_1|Z)$ becomes subadditive (or equivalently, where $\widehat{H}(X_0X_1Z)$ becomes strongly subadditive) is exactly when the PR-box is weakened to obey Tsirelson's bound. Note that our trivial example only shows that PR-boxes that are more than $\approx 0.89 > 1/2 + 1/(2\sqrt{2})$ correct do not obey subadditivity. However, note that constraining non-local boxes to obey Tsirelson's bound alone is insufficient to reduce box world to quantum theory (e.g. each quantum system admits a continuum of fine-grained measurements, whereas any box admits only a finite set).

(c) Conditioning can increase entropy. Our small example also emphasizes another curious property of the conditional entropy. By definition,

$$\widehat{H}(X_0|X_1Z) = \widehat{H}(X_0X_1Z) - \widehat{H}(X_1Z) = 1. \tag{46}$$

But this is strange, because we can perfectly determine the output of X_0 given Z. Furthermore, since $\widehat{H}(X_0|Z) = 0$, we then clearly have

$$\widehat{H}(X_0|X_1Z) > \widehat{H}(X_0|Z), \tag{47}$$

which means that 'forgetting information', namely discarding X_1, can *decrease* uncertainty. Again, it may seem that this is a consequence of not choosing the 'correct' definition of entropy.

5.2. *Alternative conditional entropy*

Reevaluating the conditional entropies of the previous section using this new definition, we find that

$$\widehat{H}_+(X_0|Z) = \widehat{H}_+(X_1|Z) = 0, \quad \widehat{H}_+(X_0X_1|Z) = 1 \tag{48}$$

as before; hence, this new measure still violates subadditivity. However, we now have

$$\widehat{H}_+(X_0|ZX_1) = 0, \tag{49}$$

as we would intuitively expect. This means that conditioning on X_1 no longer increases the entropy. However, it generates a violation of the chain rule

$$\widehat{H}_+(X_0X_1|Z) \neq \widehat{H}_+(X_1|Z) + \widehat{H}_+(X_0|X_1Z). \tag{50}$$

On balance though, this measure of conditional entropy seems more reasonable than the original one in this example.

6. Properties of conditional entropies in box world

We now show that *any* 'reasonable' measure of the conditional entropy in box world will necessarily defy our intuition about information in several ways.

Intuitively, the goal of any entropic quantity is to capture the degree of uncertainty we have about a system, possibly given access to some additional information. We assign a label A to the system of interest and use B to denote any additional systems or information available to us. For simplicity, let us suppose that A corresponds to some classical information (i.e. it is a state of a classical box). Let $\widetilde{H}(A|B)$ denote some entropic quantity that quantifies our uncertainty about A given B. If we were able to determine A with certainty given access to B (i.e. to determine the precise output of the classical box A), we would intuitively say that there is no uncertainty and the quantity $\widetilde{H}(A|B)$ should vanish. Conversely, if we cannot determine A given B, but will necessarily have some residual uncertainty, then the quantity $\widetilde{\mathrm{H}}(\mathrm{A}|\mathrm{B})$ should be positive. Motivated by this intuition in quantifying uncertainty, we demand the following two properties to hold for any 'reasonable' measure of uncertainty when A is classical.

{1} If the output of A can be obtained from B with certainty, $\widetilde{\mathrm{H}}(\mathrm{A}|\mathrm{B}) = 0$.

{2} If the output of A cannot be obtained from B with certainty, then $\widetilde{\mathrm{H}}(\mathrm{A}|\mathrm{B}) > 0$.

In the classical and quantum worlds, all commonly used entropic quantities satisfy these conditions (given that A is classical). In both such worlds, there also exist entropic quantities that are subadditive and obey a chain rule, e.g. the conditional Shannon and von Neumann entropies. In box world, $\widehat{\mathrm{H}}_+(\mathrm{A}|\mathrm{B})$ is 'reasonable' according to this definition, while $\widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B})$ is 'unreasonable'. Curiously, it turns out that in box world there cannot be *any* reasonable measure of conditional entropy that obeys conditions {1} and {2}, but at the same time is subadditive or obeys a chain rule.

(a) Subadditivity of the conditional entropy. Consider the state of the two classical bits $\mathrm{A} = \mathrm{X}_0\mathrm{X}_1$ and Bob's binary input/output box $\mathrm{B} = \mathrm{Z}$ described by (37) in the previous section. We now show that in this case *no* reasonable measure of entropy that obeys properties {1} and {2} is subadditive. First of all, note that Bob can determine one of the bits perfectly, given access to Z. Therefore from condition {1}, we obtain that

$$\widetilde{\mathrm{H}}(\mathrm{X}_0|\mathrm{Z}) = \widetilde{\mathrm{H}}(\mathrm{X}_1|\mathrm{Z}) = 0. \tag{51}$$

However, since Bob cannot determine the parity of the two bits, he certainly cannot learn both bits perfectly and hence, from condition {2}, we have

$$\widetilde{\mathrm{H}}(\mathrm{X}_0\mathrm{X}_1|\mathrm{Z}) > 0. \tag{52}$$

In order for subadditivity to hold, we would need that

$$\widetilde{\mathrm{H}}(\mathrm{X}_0\mathrm{X}_1|\mathrm{Z}) \leqslant \widetilde{\mathrm{H}}(\mathrm{X}_0|\mathrm{Z}) + \widetilde{\mathrm{H}}(\mathrm{X}_1|\mathrm{Z}), \tag{53}$$

which using (51) and (52) leads to a contradiction. Note that subadditivity could still hold if the quantity $\widetilde{\mathrm{H}}(\mathrm{X}_0\mathrm{X}_1|\mathrm{Z})$ were negative.

(b) Chain rule for the conditional entropy. We now show that a chain rule is impossible in box world for any entropic quantity that satisfies {1} and {2}. In fact, for the purposes of this proof, it is sufficient to replace {2} by the weaker assumption

{2′} If the output of A cannot be obtained from B with certainty, then $\widetilde{\mathrm{H}}(\mathrm{A}|\mathrm{B}) \neq 0$.

Note that for the state described by (37), condition {1} gives us

$$\widetilde{\mathrm{H}}(\mathrm{X}_0|\mathrm{Z}, \mathrm{X}_1) = \widetilde{\mathrm{H}}(\mathrm{X}_0|\mathrm{Z}) = 0 \tag{54}$$

because x_0 can be obtained perfectly from $\mathrm{B} = \mathrm{Z}$ or $\mathrm{B} = \mathrm{Z}\mathrm{X}_1$. A chain rule for the conditional entropy would mean that

$$\widetilde{\mathrm{H}}(\mathrm{X}_0\mathrm{X}_1|\mathrm{Z}) = \widetilde{\mathrm{H}}(\mathrm{X}_1|\mathrm{Z}) + \widetilde{\mathrm{H}}(\mathrm{X}_0|\mathrm{Z}, \mathrm{X}_1). \tag{55}$$

Using equation (54), together with equations (51) and (52), again gives us a contradiction. Note that $\widehat{\mathrm{H}}_+(\cdot|\cdot)$ obeys conditions {1} and {2}, and hence does not admit a chain rule in box world.

As $\widehat{\mathrm{H}}(\cdot|\cdot)$ satisfies a chain rule, it follows from the above that it must be 'unreasonable'. Indeed, this can be seen from the fact that $\widehat{\mathrm{H}}(\mathrm{X}_0|\mathrm{X}_1\mathrm{Z}) = 1$ despite the fact that we can perfectly determine the output of X_0 given Z and X_1, violating condition $\{1\}$. It is easy to see that if we were to drop the conditions that make an entropy 'reasonable' but simply assume that it is not subadditive, but we do enforce a chain rule, then conditioning can increase entropy.

7. Information causality

We now use our entropic quantities to investigate the game given in [26]. This task relates to 'information causality', which is expressed as the principle that 'communication of k classical bits causes information gain of at most k bits'. In [26], it is reported that this principle can be violated in box world using the following simple game (where we take $k = 1$): Alice is given two random classical bits a_0 and a_1 and Bob is given a single random bit t. Alice is allowed to send a single bit message m to Bob, after which he must output a bit b. The couple succeed in the task if $b = a_t$.

This task is clearly very similar to the non-local game considered in section 5. Indeed, any solution to the previous problem can also be used to solve this one. Alice takes the parity bit as $x = a_0 \oplus a_1$, then generates x_0 and $x_1 = x_0 \oplus x$ as before. She sends the message $m = x_0 \oplus a_0$ to Bob. Using the previous protocol, Bob generates x_t, and then outputs $b = x_t \oplus m = a_t$.

In the context of this game, 'information causality' is interpreted as meaning that

$$\mathrm{I} := \mathrm{I}(a_0; b|t = 0) + \mathrm{I}(a_1; b|t = 1) \leqslant 1. \tag{56}$$

where $\mathrm{I}(\cdot;\cdot|\cdot)$ is the classical conditional mutual information. This inequality is obeyed in quantum theory. However, given the above argument, it is clear that it can be violated in box world, as Alice and Bob can achieve $\mathrm{I} = 2$.

Let us examine why (56) fails in terms of our general entropies. We consider the state just after Bob has received the message from Alice, when she holds classical bits A_0 and A_1, and Bob holds the classical message M and his part of the PR-box Z. This state is described by

$$P(a_0 a_1\, m z_{\mathrm{out}}|z_{\mathrm{in}}) = \begin{cases} \frac{1}{8} & : \quad z_{\mathrm{out}} = a_{z_{\mathrm{in}}} \oplus m, \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{57}$$

We can compute entropies explicitly in this case as in section 5, and will obtain similar results. However, Pawlowski's work [26] also contains a proof of (56) in quantum theory based on the quantum mutual information. It is interesting to attempt to follow this proof using our general mutual information $\hat{\mathrm{I}}$ (or $\hat{\mathrm{I}}_+$) to see where it fails.

The quantum proof relies on the chain rule for quantum mutual information (which $\hat{\mathrm{I}}$ satisfies by definition)[8], positivity of the mutual information (which is true for $\hat{\mathrm{I}}$ in box world due to the subadditivity of $\hat{\mathrm{H}}$) and non-signalling (which is one of the defining features of box world). However, the crucial step is a use of the data processing inequality to deduce that

$$\hat{\mathrm{I}}(\mathrm{A}_0; \mathrm{A}_1\mathrm{MZ}) \geqslant \hat{\mathrm{I}}(\mathrm{A}_0; \mathrm{MZ}). \tag{58}$$

[8] This chain rule can be expressed as $\hat{\mathrm{I}}(\mathrm{A}; \mathrm{BC}) = \hat{\mathrm{I}}(\mathrm{A}; \mathrm{C}) + \hat{\mathrm{I}}(\mathrm{A}; \mathrm{B}|\mathrm{C})$, where $\hat{\mathrm{I}}(\mathrm{A}; \mathrm{B}|\mathrm{C}) = \widehat{\mathrm{H}}(\mathrm{A}|\mathrm{C}) - \widehat{\mathrm{H}}(\mathrm{A}|\mathrm{BC})$.

Although it is very natural that 'forgetting' A_1 can only decrease the mutual information, this inequality is violated in box world. Indeed, for the state (57), we find

$$\hat{\mathrm{I}}(A_0;\, A_1 M Z) = 0, \qquad \hat{\mathrm{I}}(A_0;\, M Z) = 1. \tag{59}$$

This is again a consequence of the *violation of strong subadditivity* for $\widehat{\mathrm{H}}$, which forms the key ingredient for explanation of why (56) can be violated in box world.

Although the violation of (56) in box world, and its validity in quantum theory, is a very interesting result, it is interesting to consider whether this really implies that communicating k bits has caused an information gain of more than k bits. From the state (57) it is easy to check that

$$\hat{\mathrm{I}}(\mathrm{A_0A_1};\, \mathrm{MZ}) = \hat{\mathrm{I}}_{+}(\mathrm{A_0A_1};\, \mathrm{MZ}) = 1 \leqslant 1, \tag{60}$$

and hence under both these measures, the total information about the composite system $\mathrm{A_0A_1}$ has only increased by one bit due to the one bit classical message. In appendix C.2 we show that in box world we indeed have that, given some arbitrary system Z held by Bob, the mutual information about a classical string A can never increase by more than the length of a classical message M that is transmitted. Furthermore, Bob can extract only one of the two bits, either A_0 or A_1, with the help of the message, as is indeed noted in [26]. It is therefore arguable that the information gain of Bob is only one bit. Perhaps 'information causality' should be restated in a clearer way that more directly represents the form of (56). For example, the principle that an m bit classical communication allows us to learn any one of at most m unknown bits.

8. A simple coding theorem

We now show that for some theories, the entropic quantity $\widehat{\mathrm{H}}(\cdot)$ has an appealing operational interpretation in capturing our ability to compress information. Here, we will only show this for theories obeying further restrictions, and it is an interesting open question how generally this interpretation applies.

8.1. Dimension and subspaces

Before we can talk about compression, we first need to clarify our notions of the size of a system. Intuitively, the size of a system should limit the amount of uncertainty we can have about it. Furthermore, to compress, we will clearly need to shrink the original state space. It is therefore helpful to define a notion of size for any subset of allowed states $\mathcal{S}_{\mathrm{T}} \subseteq \mathcal{S}$.

We refer to the size of a set of states $\mathcal{S}_{\mathrm{T}}$ as its dimension d, which we define by

$$d := \min_{\mathbf{e}\in\mathcal{E}^*} |\{r \in \mathcal{R}_{\mathbf{e}} | \exists S \in \mathcal{S}_{\mathrm{T}}, \quad e_r(S) > 0\}|. \tag{61}$$

This corresponds to eliminating all measurement outcomes that cannot occur for any state in $\mathcal{S}_{\mathrm{T}}$, and then counting the minimal number of remaining outcomes for any fine-grained measurement. It follows that $\log d \geqslant \widehat{\mathrm{H}}(S)$ for all $S \in \mathcal{S}_{\mathrm{T}}$. In quantum theory, d corresponds precisely to the dimension of a Hilbert space.

A natural way to select a subset of states is to consider all states that yield a given measurement outcome with certainty. We refer to an effect f such that $\{f, u - f\}$ is an allowed measurement, and that occurs with certainty for some state, as a *full* effect (i.e. f is full if there exists $S \in \mathcal{S}$ such that $f(S) = 1$). For any full effect f, we can therefore define a non-empty subset of states $\mathcal{S}_f = \{S|S \in \mathcal{S}, f(S) = 1\}$. We refer to such a subset as the *subspace* of $\mathcal{S}$ given by f. Note that subspaces are always convex, and the subspace corresponding to an effect f that is both full *and* fine-grained obeys $d_f = 1$.

We say that we have compressed a state if we have constrained it to lie within a set of states of smaller dimension.

8.2. Additional assumptions

So far, we have not been concerned about what happens to a state after a measurement. In our compression protocol, however, we will need to use an abstract notion of post-measurement states, as described in section 2.3. In particular, we will consider *pseudo-projective measurements*, which we define to be measurements that fulfill two conditions.

(a) *Repeatability.* A pseudo-projective measurement is repeatable, such that if the same measurement is applied again the same result is obtained. This requires that the output state S_r after obtaining a result r lies in the subspace given by e_r (i.e., $e_r(S_r) = 1$). Consequently, all effects in a pseudo-projective measurement must be full effects.

(b) *Weak Disturbance.* If a particular outcome r of a pseudo-projective measurement occurs with probability $e_r(S) \geqslant 1 - \delta$ for a state S, then the post measurement state S_r after this result is obtained satisfies $e_r(S)\mathcal{D}(S, S_r) \leqslant c\delta^{\varepsilon}$, where $c \geqslant 0$ and $\varepsilon \in (0, 1]$ are constants depending on the particular theory. For example, for projective measurements in quantum theory $c = (\sqrt{8} + 1)/2$ and $\varepsilon = 1/2$.

Any projective measurement in quantum theory fulfills these conditions, but these conditions alone do not define projective measurements, hence the slightly different name. In quantum theory, the weak disturbance property can be understood as an instance of the gentle measurement lemma [43].

Furthermore, in order to prove our simple coding theorem, we will need to make some additional assumptions on the states and the measurements that achieve the minimal output entropy $\widehat{\mathrm{H}}(\cdot)$ in our theory. In particular, we assume that for all states, the minimal output entropy can be attained by a pseudo-projective measurement. That is, we assume that for all $S \in \mathcal{S}$ there exists some pseudo-projective measurement $\mathbf{e} \in \mathcal{E}^*$ such that $\widehat{\mathrm{H}}(S) = H(\mathbf{e}(S))$. We further assume that for all such measurements, $\mathbf{e}^{\otimes n}$ is fine-grained and pseudo-projective, and that course grainings of $\mathbf{e}^{\otimes n}$ can also be made pseudo-projective. Lastly, we assume that the dimension of $\mathcal{S}^{\otimes n}$ is d^n. These assumptions are all true in the classical and quantum case (where $\mathbf{e}$ is projective).

We will see in appendix E that this is all we will need to show the following simple coding theorem following the steps taken by Shannon [31] and Schumacher [30] (see for example [24]).

8.3. Compression

We consider a source that emits a state $\tilde{S}_k \in \mathcal{S}$ with probability q_k, chosen independently at random in each time step. When considering n time steps, we hence obtain a sequence of states $\tilde{S}_{\vec{k}} = \tilde{S}_{k_1}, \ldots, \tilde{S}_{k_n} \in \mathcal{S}^{\otimes n}$ with $\vec{k} = (k_1, \ldots, k_n)$, where each sequence occurs with probability $q_{\vec{k}} = \Pi_j q_{k_j}$. A compression scheme consists of an encoding and decoding procedure. The encoding procedure maps each possible $\tilde{S}_{\vec{k}}$ into a state $\hat{S}_{\vec{k}} \in \mathcal{S}_f \subset \mathcal{S}^{\otimes n}$. In turn, the decoding procedure maps the states $\hat{S}_{\vec{k}}$ back to states $\check{S}_{\vec{k}} \in \mathcal{S}$ on the original state space. In analogy with the quantum case, we say that the compression scheme has rate R, if the dimension of the smaller space obeys $d_{\mathrm{f}} \leqslant 2^{nR}$. Note that in order for a compression scheme to be useful, it must have $R < \log d$ (and hence $d_{\mathrm{f}} < d^n$). A compression scheme is called *reliable*, if we can recover the original state (almost) perfectly, in the sense that the average distance between the original and the reconstructed state can be made arbitrarily small for sufficiently large n; that is, for any $\epsilon > 0$ and all sufficiently large n,

$$\sum_k q_{\vec{k}} \mathcal{D}(\tilde{S}_{\vec{k}}, \breve{S}_{\vec{k}}) \leqslant \epsilon. \tag{62}$$

Note that the output of the source can be described as a mixed state $\mathsf{Src} = \sum_k q_k \tilde{S}_k$ in each time step, and a product state $\mathsf{Src}^{\otimes n} \in \mathcal{S}^{\otimes n}$ over the course of n time steps. We then obtain the following theorem (see appendix E) in terms of the entropy of the source $\widehat{\mathrm{H}}(\mathsf{Src})$.

Theorem 8.1. *Consider an i.i.d. source* $\{q_k, \tilde{S}_k \in \mathcal{S}\}_k$ *with entropy rate* $\widehat{\mathrm{H}}(\mathsf{Src})$. *Then for* $R > \widehat{\mathrm{H}}(\mathsf{Src})$ *there exists a reliable compression scheme with rate* R.

Note that in order to establish that $\widehat{\mathrm{H}}(\cdot)$ truly characterizes our ability to compress information, we would also like to have a converse stating that for $R < \widehat{\mathrm{H}}(\mathsf{Src})$ there exists no reliable compression scheme. In quantum theory, it is not hard to prove the converse of the above theorem since it admits a strong duality between states and measurements, which may also hold for other theories. Here, however, we explicitly tried to avoid introducing any such strong assumptions.

9. Conclusion and open questions

We introduced entropic measures to quantify information in any physical theory that admits minimal notions of systems, states and measurements. Even though these measures necessarily have some limitations, we nevertheless showed that they also exhibit many intuitive properties, and for some theories have an appealing operational interpretation, quantifying our ability to compress states. Most of the problems we encountered with the conditional entropy seem to arise due to a violation of strong subadditivity. It is an interesting question whether quantum and classical theories are the only ones in which $\widehat{\mathrm{H}}$ is strongly subadditive, or whether this is true for other theories. Indeed, it would be an exciting question to turn things around and start by demanding that our entropic measures *do* satisfy these properties, and determine how this restricts the set of possible theories.

In $\widehat{\mathrm{H}}_{+}(\cdot|\cdot)$, we defined a natural entropic quantity that differs from the conditional von Neumann entropy in quantum theory and has been used in [12] to study quantum correlations. It would be interesting to study whether this quantity can shed any further light on quantum phenomena, or if an alternative conditional entropy can be defined that behaves like $\widehat{\mathrm{H}}_{+}(\cdot|\cdot)$ in box world, but still reduces to the conditional Shannon entropy in quantum theory.

While we have proved some intuitive properties of our quantities, it would be interesting to see whether other properties of the von Neumann or Shannon entropies carry over to this setting. In particular, it would be interesting to prove bounds on the mutual and accessible information analogous to Holevo's theorem when none of the systems are classical.

Another interesting question is whether one can find a closed form expression for the relative entropy in general theories. In quantum theory, we can define the mutual information (and indeed the entropy itself) in terms of the relative entropy[9]; hence, such an approach may also yield an alternative definition of other entropic quantities for general theories.

We believe our measures are an interesting step toward understanding information processing in general physical theories, which may in turn shed some light on our own quantum world.

Acknowledgments

The non-local game used in our example above was discovered in collaboration with Andrew Doherty, whom we thank for his kind permission to use it here. The authors also thank Sergio Boixo, Matthew Elliot and Jonathan Oppenheim for interesting discussions, and Matt Leifer and Ronald de Wolf for comments on an earlier draft. SW is supported by NSF grants PHY-04056720 and PHY-0803371. AJS is supported by a Royal Society URF, and in part by the EU QAP project (CT-015848). Part of this work was done while AJS was visiting Caltech (Pasadena, USA).

Note added. During the course of this work, we learned that an independent work on the same general topic [40] is to appear in *New Journal of Physics*. Yet another related work appeared subsequently [41].

Appendix A. Distance metric

In this appendix, we provide formal statements and the technical details of our claims.

We now show that the quantity (8) is indeed a metric on the state space $\mathcal{S}$.

Lemma A.1. $\mathcal{D}: \mathcal{S}\times\mathcal{S} \to [0,1]$ *as defined in (8) is a metric on the state space* $\mathcal{S}$.

Proof. Consider states $S_0, S_1, S_2 \in \mathcal{S}$. Clearly,

$$\mathcal{D}(S_0, S_1) \geqslant 0 \tag{A.1}$$

using the property of the classical statistical distance, where equality holds iff $S_0 = S_1$ by definition of the state space $\mathcal{S}$. It remains to show that $\mathcal{D}$ obeys a triangle inequality. Let $\mathbf{e}_{ij}$ be the optimal measurement to distinguish states i and j. We then have

$$\begin{aligned}\mathcal{D}(S_0, S_1) + \mathcal{D}(S_1, S_2) &\geqslant \mathcal{C}(\mathbf{e}_{02}(S_0), \mathbf{e}_{02}(S_1)) + \mathcal{C}(\mathbf{e}_{02}(S_1), \mathbf{e}_{02}(S_2))\\ &\geqslant \mathcal{C}(\mathbf{e}_{02}(S_0), \mathbf{e}_{02}(S_2)) = \mathcal{D}(S_0, S_2),\end{aligned} \tag{A.2}$$

where the second inequality follows from the fact that the classical statistical distance C itself obeys the triangle inequality. □

Appendix B. Properties of $\widehat{\mathrm{H}}$

In this appendix, we derive the properties of the entropy $\widehat{\mathrm{H}}$ used in the paper. Note that by assumption $\mathcal{E}^*$ is non-empty, which implies that $\widehat{\mathrm{H}}(S)$ is well defined.

[9] In particular, the mutual information for a quantum state ρ_{AB} is the same as the relative entropy between ρ_{AB} and $\rho_{\mathrm{A}} \otimes \rho_{\mathrm{B}}$, and the entropy of ρ is (minus) the relative entropy between ρ and the identity operator.

B.1. Reduction to the von Neumann and the Shannon entropies

We now show that the entropic quantity (18) reduces to the von Neumann and the Shannon entropies in the classical and quantum settings, respectively. For the relation to the von Neumann entropy, we will need the following little lemma.

Lemma B.1. *Let $\rho \in \mathcal{B}(\mathcal{H})$ be a quantum state with eigendecomposition $\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|$. Then*

$$\widehat{\mathrm{H}}(\rho) = \mathrm{S}(\rho) = \mathrm{H}(\vec{p}), \tag{B.1}$$

where $\vec{p} = (p_1, \ldots, p_d)$ with $d = \dim(\mathcal{H})$.

Proof. Our goal will be to show that for any fine-grained measurement $\mathbf{e}$ with

$$e_l = c_\ell |\phi_\ell\rangle\langle\phi_\ell| \in \mathcal{B}(\mathcal{H}) \ \mid\ 0 \leqslant c_\ell \leqslant 1, \quad \text{and} \quad \sum_\ell c_\ell |\phi_\ell\rangle\langle\phi_\ell| = \mathbb{I}, \tag{B.2}$$

the Shannon entropy of the distribution $q_\ell := c_\ell \langle\phi_\ell|\rho|\phi_\ell\rangle$ is always at least as large as the distribution obtained by measuring in the eigenbasis of ρ, that is,

$$\mathrm{H}(\vec{p}) \leqslant \mathrm{H}(\vec{q}), \tag{B.3}$$

with $\vec{q} = (q_1, \ldots, q_N)$.

Let $N = |\mathbf{e}|$ and note that $d \leqslant N$. First, of all note that we can always extend a distribution $\{p_j\}$ over d elements to a distribution $\{\tilde{p}_j\}$ over N elements by letting $\tilde{p}_j = p_j$ for all $j \leqslant d$ and $\tilde{p}_j = 0$ for all $j > d$. Clearly, $\mathrm{H}(\vec{\tilde{p}}) = \mathrm{H}(\vec{p})$ with $\vec{\tilde{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N)$.

Secondly, note that

$$q_\ell = \sum_j p_j q_{\ell|j} \qquad \text{and} \qquad q_{\ell|j} = c_\ell |\langle\phi_\ell|\psi_j\rangle|^2, \tag{B.4}$$

from which we immediately obtain together with (B.2) that

$$\sum_j q_{\ell|j} = c_\ell \qquad \text{and} \qquad \sum_\ell q_{\ell|j} = 1. \tag{B.5}$$

Consider the $N \times N$ matrix M determined by the entries

$$M_{\ell,j} = \begin{cases} q_{\ell|j}, & \text{for } j \leqslant d, \\ \dfrac{1 - c_\ell}{N - d}, & \text{for } j > d. \end{cases} \tag{B.6}$$

which allows us to write $\vec{q} = M\vec{\tilde{p}}$. Note that since $M_{\ell,j} \geqslant 0$ and $\sum_j M_{\ell,j} = \sum_\ell M_{\ell,j} = 1$, M is a doubly stochastic matrix. Using Birkhoff's theorem (see e.g. [19, theorem 8.7.1]), we may thus write M as a convex combination of permutation matrices, that is,

$$M = \sum_{\pi \in S_N} P(\pi)\pi, \tag{B.7}$$

where P is a probability distribution over the group of permutations S_N. Using the concavity of the Shannon entropy, we obtain

$$\mathrm{H}(\vec{q}) \geqslant \sum_{\pi \in S_N} P(\pi)\mathrm{H}(\pi(\vec{p})) = \mathrm{H}(\vec{p}). \tag{B.8}$$

As we can always measure ρ in its eigenbasis, it follows that

$$\widehat{\mathrm{H}}(\rho) = \inf_{\mathbf{e}\in\mathcal{E}^*} \mathrm{H}(\mathbf{e}(S)) = \inf_{\vec{q}} \mathrm{H}(\vec{q}) = \mathrm{H}(\vec{p}) \tag{B.9}$$

and it is easy to see that $\mathrm{H}(\vec{p}) = \mathrm{S}(\rho)$. □

Since the von Neumann entropy reduces to the Shannon entropy in a classical setting, this also shows that the entropic quantity (18) reduces to the Shannon entropy in the classical case.

B.2. Positivity, boundedness, and concavity

Here we prove the other general properties of the entropy $\widehat{\mathrm{H}}$.

Positivity. This follows trivially from the positivity of the Shannon entropy.

Boundedness. The existence of a measurement $\mathbf{e} \in \mathcal{E}^*$ with d outcomes, combined with the fact that the Shannon entropy is maximized for a uniform probability distribution, ensures that

$$\widehat{\mathrm{H}}(S) \leqslant \mathrm{H}(\mathbf{e}(S)) \leqslant \log(d), \tag{B.10}$$

which gives boundedness.

Concavity. To see that $\widehat{\mathrm{H}}$ is concave, suppose first that the infimum in the definition (18) of $\widehat{\mathrm{H}}(S_{\mathrm{mix}})$ is achieved, such that $\widehat{\mathrm{H}}(S_{\mathrm{mix}}) = \mathrm{H}(\mathbf{e}(S_{\mathrm{mix}}))$ for some $\mathbf{e} \in \mathcal{E}^*$. As effects are linear maps, $\mathbf{e}(S_{\mathrm{mix}}) = p\mathbf{e}(S_1) + (1-p)\mathbf{e}(S_2)$. Hence, by the concavity of the Shannon entropy

$$\begin{aligned}\widehat{\mathrm{H}}(S_{\mathrm{mix}}) &= \mathrm{H}(\mathbf{e}(S_{\mathrm{mix}}))\\ &\geqslant p\mathrm{H}(\mathbf{e}(S_1)) + (1-p)\mathrm{H}(\mathbf{e}(S_2))\\ &\geqslant p\widehat{\mathrm{H}}(S_1) + (1-p)\widehat{\mathrm{H}}(S_2),\end{aligned} \tag{B.11}$$

which concludes our claim. On the other hand, if the infimum is not achievable then for all sufficiently small $\delta > 0$, we can find an $\mathbf{e} \in \mathcal{E}^*$ such that $\widehat{\mathrm{H}}(S_{\mathrm{mix}}) = \mathrm{H}(\mathbf{e}(S_{\mathrm{mix}})) - \delta$. Using the same argument as before, we find

$$\widehat{\mathrm{H}}(S_{\mathrm{mix}}) \geqslant p\widehat{\mathrm{H}}(S_1) + (1-p)\widehat{\mathrm{H}}(S_2) - \delta. \tag{B.12}$$

As this holds for all sufficiently small δ, the result follows.

B.3. Limited subadditivity and continuity

Here we prove two properties of $\widehat{\mathrm{H}}$ that require additional minor assumptions on our theory. However, they are obeyed in quantum theory, classical theory and box world.

Limited subadditivity. Given an additional reasonable assumption, we can prove that $\widehat{\mathrm{H}}$ is subadditive, we first assume that there exist $\mathbf{e} \in \mathcal{E}^*_{\mathrm{A}}$ and $\mathbf{f} \in \mathcal{E}^*_{\mathrm{B}}$ such that $\widehat{\mathrm{H}}(\mathrm{A}) = \mathrm{H}(\mathbf{e}(\mathrm{A}))$ and

$\widehat{\mathrm{H}}(\mathrm{B}) = \mathrm{H}(\mathbf{f}(\mathrm{B}))$. By assumption, $\mathbf{e} \otimes \mathbf{f}$ is a fine-grained measurement on the joint system $\mathcal{AB}$. Thus, by the subadditivity of the Shannon entropy

$$\begin{aligned}\widehat{\mathrm{H}}(\mathrm{A}) + \widehat{\mathrm{H}}(\mathrm{B}) &= \mathrm{H}(\mathbf{e}(\mathrm{A})) + \mathrm{H}(\mathbf{f}(\mathrm{B})) \\ &\geqslant \mathrm{H}((\mathbf{e} \otimes \mathbf{f})\mathrm{AB}) \\ &\geqslant \widehat{\mathrm{H}}(\mathrm{AB}),\end{aligned} \tag{B.13}$$

as claimed. Now suppose that the infimum for one or both of $\widehat{\mathrm{H}}(\mathrm{A})$ or $\widehat{\mathrm{H}}(\mathrm{B})$ is not achieved. Then, for all sufficiently small $\delta > 0$, we can find $\mathbf{e} \in \mathcal{E}^*_{\mathrm{A}}$ and $\mathbf{f} \in \mathcal{E}^*_{\mathrm{B}}$ such that

$$\widehat{\mathrm{H}}(\mathrm{A}) + \widehat{\mathrm{H}}(\mathrm{B}) = \mathrm{H}(\mathbf{e}(\mathrm{A})) + \mathrm{H}(\mathbf{f}(\mathrm{B})) - \delta \geqslant \widehat{\mathrm{H}}(\mathrm{AB}) - \delta. \tag{B.14}$$

As this holds for all sufficiently small $\delta > 0$, the result follows.

Note that if A and B are in a product state, and the theory only allows product measurements on AB, then equality holds in (22). However, given that we allow an arbitrary set of joint measurements, equality does not hold when A and B are in a product state for any possible probabilistic theories (consider the case where $\widehat{\mathrm{H}}(\mathrm{A}) > \log 2$, but there exists a fine-grained measurement on AB with only two outcomes).

Limited continuity. Here we prove an analogue of the Fannes inequality [14], given an additional reasonable assumption that we can restrict to measurements with at most D outcomes without changing the entropy of a system.

Suppose without loss of generality that $\widehat{\mathrm{H}}(S_1) \geqslant \widehat{\mathrm{H}}(S_2)$. Initially, we also suppose that the infimum in the definition of $\widehat{\mathrm{H}}(S_2)$ is achieved for some $\mathbf{f} \in \mathcal{E}^*$, such that $\widehat{\mathrm{H}}(S_2) = \mathrm{H}(\mathbf{f}(S_2))$. We can then bound

$$\begin{aligned}|\widehat{\mathrm{H}}(S_1) - \widehat{\mathrm{H}}(S_2)| &\leqslant |\mathrm{H}(\mathbf{f}(S_1)) - \mathrm{H}(\mathbf{f}(S_2))| \\ &\leqslant \mathcal{C}(\mathbf{f}(S_1), \mathbf{f}(S_2)) \log\left(\frac{D}{\mathcal{C}(\mathbf{f}(S_1), \mathbf{f}(S_2))}\right) \\ &\leqslant \mathcal{D}(S_1, S_2) \log\left(\frac{D}{\mathcal{D}(S_1, S_2)}\right),\end{aligned} \tag{B.15}$$

where the first inequality follows from the fact that $\widehat{\mathrm{H}}(S_1) \leqslant \mathrm{H}(\mathbf{f}(S_1))$, the second inequality from the Fannes inequality [14] applied to the classical case and the final inequality by noting that

$$\mathcal{C}(\mathbf{f}(S_1), \mathbf{f}(S_2)) \leqslant \mathcal{D}(S_1, S_2) < \frac{1}{e}, \tag{B.16}$$

If the infimum is not achieved, then for all sufficiently small $\delta > 0$ there nevertheless exists $\mathbf{f} \in \mathcal{E}^*$ such that $\widehat{\mathrm{H}}(S_2) = \mathrm{H}(\mathbf{f}(S_2)) - \delta$. Following the same procedure as before, we find

$$|\widehat{\mathrm{H}}(S_1) - \widehat{\mathrm{H}}(S_2)| \leqslant \mathcal{D}(S_1, S_2) \log\left(\frac{D}{\mathcal{D}(S_1, S_2)}\right) + \delta \tag{B.17}$$

from which the result follows.

Appendix C. Properties of the conditional entropy

C.1. General case

We now show that in contrast to the quantity $\widehat{\mathrm{H}}$, our second form of conditional entropy $\widehat{\mathrm{H}}_{+}$ obeys the intuitive property that conditioning reduces entropy in all cases.

Lemma C.1. (Conditioning reduces entropy for $\widehat{\mathrm{H}}_{+}$). *For any tripartite state $ABC \in \mathcal{S}_{ABC}$ and its corresponding reduced states, we have*

$$\widehat{\mathrm{H}}_{+}(A) \geqslant \widehat{\mathrm{H}}_{+}(A|B) \geqslant \widehat{\mathrm{H}}_{+}(A|BC). \tag{C.1}$$

Proof. The first inequality follows by choosing the unit measurement in the infimum over $\mathcal{E}_{\mathrm{B}}$ in the definition of $\widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{B})$, and noting that $\widehat{\mathrm{H}}(\mathrm{A}) = u(\mathrm{B})\widehat{\mathrm{H}}(\mathrm{A}_{|u}) \geqslant \widehat{\mathrm{H}}(\mathrm{A}|\mathrm{B})$. The second inequality comes from restricting to measurements of the form $\mathbf{f}_{\mathrm{B}} \otimes \mathbf{u}_{C}$ in the infimum over $\mathcal{E}_{\mathrm{BC}}$ in the definition of $\widehat{\mathrm{H}}_{+}(\mathrm{A}|\mathrm{BC})$. □

C.2. Box world

We now prove a very restricted form of chain rule in box world. This will allow us to show that for our notions of entropy, the mutual information about any classical information given an arbitrary state in box world can never increase by more than ℓ bits when transmitting ℓ bits of information. To show our simple chain rule, we will use the fact that in box world, we have that when considering a composite of a classical system M and an arbitrary system B, the only allowed measurements on the composite system MB take the form of first performing the only allowed measurement on M, followed by a choice of measurement on B that may depend on the outcome of the measurement on M. Since classical systems in box world admit exactly one measurement (possibly followed by some classical post-processing), we simply write $\mathrm{H}(\mathrm{M}) = \widehat{\mathrm{H}}(\mathrm{M})$ to denote the resulting entropy.

Lemma C.2. (Box chain rule). *For any tripartite state $CMB \in \mathcal{S}_{CMB}$ in box world, where its corresponding reduced states C and M are classical, we have*

$$\widehat{\mathrm{H}}_{+}(C|MB) \geqslant \widehat{\mathrm{H}}_{+}(CM|B) - \widehat{\mathrm{H}}(M). \tag{C.2}$$

Proof. For simplicity, we only examine the case where the infimum is attained in $\widehat{\mathrm{H}}_{+}$, and the other case can again be obtained by taking the appropriate limit. Since the only measurements on MB are as described above, we clearly have

$$\begin{aligned}
\widehat{\mathrm{H}}_{+}(\mathrm{C}|\mathrm{MB}) &= \sum_{m} e_m(\mathrm{M}) \sum_{k} f_k(\mathrm{B}_{|m})\widehat{\mathrm{H}}_{+}(\mathrm{C}_{|m,k}) \\
&= \sum_{m} e_m(M) \sum_{k} f_k(\mathrm{B}_{|m})\mathrm{H}(\mathrm{C}|\mathrm{M} = m, K = k) \\
&= \mathrm{H}(\mathrm{C}|\mathrm{M}, K) \\
&= \mathrm{H}(\mathrm{CM}|K) - \mathrm{H}(\mathrm{M}|K) \\
&\geqslant \widehat{\mathrm{H}}_{+}(\mathrm{CM}|\mathrm{B}) - \widehat{\mathrm{H}}(\mathrm{M},
\end{aligned} \tag{C.3}$$

where the first equality follows from the definition of $\widehat{\mathrm{H}}_+$ and the fact that M is classical, the second from the definition of the conditional Shannon entropy, the third from the chain rule for the conditional Shannon entropy, and the final inequality from the definition of $\widehat{\mathrm{H}}_+$, the fact that $\widehat{\mathrm{H}}(\mathrm{M}) = \mathrm{H}(\mathbf{e}(\mathrm{M}))$ for classical systems and the fact that conditioning reduces entropy for the Shannon entropy. □

We now see that consistent with the no-signalling principle, the transmission of an ℓ bit message M causes the mutual information about a classical system C given access to some arbitrary box information B to increase by at most ℓ bits. Note that for our alternate definition of conditional entropy and mutual information, we have

$$\hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{MB}) = \widehat{\mathrm{H}}(\mathrm{C}) - \widehat{\mathrm{H}}_+(\mathrm{C}|\mathrm{MB}). \tag{C.4}$$

First, note that we can write

$$\begin{aligned} \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{MB}) &= \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{B}) + \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{M}|\mathrm{B}),\\ \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{B}) &= \widehat{\mathrm{H}}(\mathrm{C}) - \widehat{\mathrm{H}}_+(\mathrm{C}|\mathrm{B}),\\ \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{M}|\mathrm{B}) &:= \widehat{\mathrm{H}}_+(\mathrm{C}|\mathrm{B}) - \widehat{\mathrm{H}}_+(\mathrm{C}|\mathrm{MB}), \end{aligned} \tag{C.5}$$

by definition. Hence, we have

$$\begin{aligned} \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{MB}) &\leqslant \widehat{\mathrm{H}}_+(\mathrm{C}|\mathrm{B}) + \widehat{\mathrm{H}}_+(\mathrm{M}) - \widehat{\mathrm{H}}_+(\mathrm{CM}|\mathrm{B})\\ &\leqslant \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{B}) + \widehat{\mathrm{H}}_+(\mathrm{M}) \leqslant \hat{\mathrm{I}}_+(\mathrm{C};\,\mathrm{B}) + \ell. \end{aligned} \tag{C.6}$$

Appendix D. Properties of $\breve{\mathrm{H}}$

In this section, we explore properties of the decomposition entropy $\breve{\mathrm{H}}$.

D.1. Reduction to the von Neumann and the Shannon entropies

To show the reduction of $\breve{\mathrm{H}}(\rho)$ to the von Neumann entropy $S(\rho)$ in quantum theory, we use the following lemma.

Lemma D.1 (theorem 11.10 in [24]). *Suppose that $\rho = \sum_i p_i \rho_i$, where p_i are some set of probabilities and ρ_i are density operators. Then*

$$\mathrm{S}(\rho) \leqslant \sum_i p_i \mathrm{S}(\rho_i) + \mathrm{H}(p_i), \tag{D.1}$$

with equality if and only if the states ρ_i have support on orthogonal subspaces.

Note that when ρ_i are pure states, $\mathrm{S}(\rho_i) = 0$. Hence for any pure state decomposition $\mathbf{D}(\rho)$, this implies

$$\mathrm{S}(\rho) \leqslant \mathrm{H}(\mathbf{D}(\rho)). \tag{D.2}$$

Furthermore, denoting an eigendecomposition of ρ by $\mathbf{D}^*(\rho)$, it is easy to see that $\mathrm{H}(\mathbf{D}^*(\rho)) = \mathrm{S}(\rho)$. Hence, it follows that

$$\mathrm{S}(\rho) = \breve{\mathrm{H}}(\rho) = \inf_{\mathbf{D}(\rho)} \mathrm{H}(\mathbf{D}(\rho)). \tag{D.3}$$

D.2. Subadditivity and concavity

In this section, we will show that $\breve{\mathrm{H}}$ is neither concave nor subadditive by giving explicit counterexamples from box world.

First consider a single box with binary input/output. For clarity, we will represent its state by giving its probability distribution $P(a|x)$ in vector form:

$$S = \begin{pmatrix} P(0|0) \\ P(1|0) \\ \hline P(0|1) \\ P(1|1) \end{pmatrix}. \tag{D.4}$$

Now consider the two states

$$S_1 = \begin{pmatrix} 1 \\ 0 \\ \hline 1/2 \\ 1/2 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 1/2 \\ 1/2 \\ \hline 1 \\ 0 \end{pmatrix}, \tag{D.5}$$

These can both be optimally decomposed into two equally weighted pure states, e.g.

$$S_1 = \frac{1}{2}\begin{pmatrix} 1 \\ 0 \\ \hline 1 \\ 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ 0 \\ \hline 0 \\ 1 \end{pmatrix} \tag{D.6}$$

and hence they satisfy $\breve{\mathrm{H}}(S_1) = \breve{\mathrm{H}}(S_2) = \log 2 = 1$. However, now consider the mixed state,

$$S_{\mathrm{mix}} = \frac{1}{2}S_1 + \frac{1}{2}S_2 = \begin{pmatrix} 3/4 \\ 1/4 \\ \hline 3/4 \\ 1/4 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 \\ 0 \\ \hline 1 \\ 0 \end{pmatrix} + \frac{3}{4}\begin{pmatrix} 0 \\ 1 \\ \hline 0 \\ 1 \end{pmatrix}, \tag{D.7}$$

which has $\breve{\mathrm{H}}(S_{\mathrm{mix}}) = \mathrm{H}((\frac{3}{4}, \frac{1}{4})) < 1$. Hence, in this case, we violate concavity

$$\breve{\mathrm{H}}(S_{\mathrm{mix}}) < \tfrac{1}{2}\breve{\mathrm{H}}(S_1) + \tfrac{1}{2}\breve{\mathrm{H}}(S_2). \tag{D.8}$$

To obtain a violation of subadditivity, we consider a bipartite state in which each system has a binary input/output, represented in the form of a matrix

$$S_{\mathrm{AB}} = \left(\begin{array}{cc|cc} P(00|00) & P(01|00) & P(00|01) & P(01|01) \\ P(10|00) & P(11|00) & P(10|01) & P(11|01) \\ \hline P(00|10) & P(01|10) & P(00|11) & P(01|11) \\ P(10|10) & P(11|10) & P(10|11) & P(11|11) \end{array}\right). \tag{D.9}$$

Choose the following allowed state

$$S_{\mathrm{AB}} = \frac{1}{8}\left(\begin{array}{cc|cc} 2 & 3 & 2 & 3 \\ 3 & 0 & 3 & 0 \\ \hline 5 & 0 & 2 & 3 \\ 0 & 3 & 3 & 0 \end{array}\right), \qquad A = B = \frac{1}{8}\begin{pmatrix} 5 \\ 3 \\ \hline 5 \\ 3 \end{pmatrix}. \tag{D.10}$$

It is known that in this case there are exactly 24 pure states for the bipartite binary input/output case (16 product states and eight entangled states) [3], which we denote by S^i_{AB}. By demanding that $S_{\mathrm{AB}} - p_i S^i_{\mathrm{AB}}$ be a positive matrix for each pure state, we find that any decomposition must satisfy $p_i \leqslant \frac{1}{4}\,\forall\, i$. Hence, $\breve{\mathrm{H}}(\mathrm{AB}) = \inf_{\mathbf{D}(\rho)} \mathrm{H}(p_i) \geqslant 2$. In fact, we can construct an explicit decomposition in terms of an entangled state and three product states (all equally weighted), giving $\breve{\mathrm{H}}(\mathrm{AB}) = 2$. The marginal states on the other hand, satisfy

$$\breve{\mathrm{H}}(\mathrm{A}) = \breve{\mathrm{H}}(\mathrm{B}) = \mathrm{H}\left(\left(\frac{3}{8}, \frac{5}{8}\right)\right) < 1. \tag{D.11}$$

Hence, we obtain

$$\breve{\mathrm{H}}(\mathrm{AB}) > \breve{\mathrm{H}}(\mathrm{A}) + \breve{\mathrm{H}}(\mathrm{B}) \tag{D.12}$$

in violation of subadditivity.

Appendix E. A simple coding theorem

We now sketch the proof of theorem 8.1, which is straightforward following the steps taken in the quantum setting [30].

Consider the pseudo-projective measurement $\mathbf{e}$ that gives the minimal output entropy for the state Src, which we take to exist by assumption.

At the core of our little coding theorem lies an observation about ε-typical sequences analogous to the classical and quantum setting. Define the set of ε-typical outcomes when measuring $\mathbf{e}^{\otimes n}$ on the state $\mathsf{Src}^{\otimes n} \in \mathcal{S}^{\otimes n}$ as

$$T(n,\varepsilon) := \left\{ r_1, \ldots, r_n \in \mathcal{R}_{\mathbf{e}}^{\times n} \mid \left| \frac{1}{n} \log\left(\frac{1}{e_{r_1}(\mathsf{Src}) \ldots e_{r_n}(\mathsf{Src})} \right) - \widehat{\mathrm{H}}(\mathsf{Src}) \right| \leqslant \varepsilon \right\}. \tag{E.1}$$

When n and ε are clear from context, we will also use the effects

$$h_{\mathrm{T}} := \sum_{\vec{r} \in T(n,\varepsilon)} e_{\vec{r}}, \tag{E.2}$$

$$h_{\mathrm{A}} := u - h_{\mathrm{T}}. \tag{E.3}$$

Since we assumed that any theory contains arbitrary coarse-grainings of measurements, we can consider the measurement

$$\mathbf{h} := \{(T, h_{\mathrm{T}}), (A, h_{\mathrm{A}})\}, \tag{E.4}$$

which by assumption we can make pseudo-projective. We refer to the subspaces given by h_{T} and h_{A} as the typical and atypical subspaces, respectively. If we observe outcome 'T' for the measurement $\mathbf{h}$, we conclude that a state lies in the typical subspace associated with the set $T(n,\varepsilon)$. Otherwise, we conclude that the states lies in the atypical subspace.

Note that by assumption, we have that $\mathbf{e}^{\otimes n}$ is a fine-grained measurement. For all states in the typical subspace, only outcomes in the typical set $T(n,\varepsilon)$ will occur. Hence, we have that the dimension of the typical subspace satisfies $d_{\mathrm{T}} \leqslant |T(n,\varepsilon)|$.

We are now ready to prove the following theorem:

Theorem E.1 (typical subspace theorem). *Let all quantities be defined as above. Fix* $\varepsilon > 0$, *then for any* $\delta > 0$ *and sufficiently large* n,

$$(i)\ \ h_{\mathrm{T}}(\mathsf{Src}^{\otimes n}) \geqslant 1-\delta. \tag{E.5}$$

$$(ii)\ \ (1-\delta)2^{n(\widehat{\mathrm{H}}(\mathsf{Src})-\varepsilon)} \leqslant |T(n,\varepsilon)| \leqslant 2^{n(\widehat{\mathrm{H}}(\mathsf{Src})+\varepsilon)}. \tag{E.6}$$

Proof. The proof of (i) and (ii) is analogous to [24, theorem 12.5] by noting that

$$h_{\mathrm{T}}(\mathsf{Src}^{\otimes n}) = \sum_{(r_1,\ldots,r_n)\in T(n,\varepsilon)} e_{r_1}(\mathsf{Src})e_{r_2}(\mathsf{Src})\ldots e_{r_n}(\mathsf{Src}), \tag{E.7}$$

and that the condition characterizing the set $T(n,\varepsilon)$ of ε-typical sequences can also be written as

$$2^{-n(\widehat{\mathrm{H}}(\mathsf{Src})+\varepsilon)} \leqslant e_{r_1}(\mathsf{Src})\ldots e_{r_n}(\mathsf{Src}) \leqslant 2^{-n(\widehat{\mathrm{H}}(\mathsf{Src})-\varepsilon)}. \tag{E.8}$$

Given the statement about typical sequences, we can now complete the proof of theorem 8.1: recall that the source emits a sequence of states $\tilde{S}_{\vec{k}}$ with probability $q_{\vec{k}}$. To compress the state, we perform a pseudo-projective measurement of $\mathbf{h}$ given by (E.4). If we obtain outcome 'T' (corresponding to the typical subspace), we output the post-measurement state $T[\tilde{S}_{\vec{k}}]$, which must lie in the typical subspace as the measurement is repeatable. Otherwise, we prepare an arbitrary fixed state in the typical subspace which we will call S_{fail}. The resulting state is thus a mixed state in the typical subspace of the form

$$\hat{S}_{\vec{k}} = h_{\mathrm{T}}(\tilde{S}_{\vec{k}})T[\tilde{S}_{\vec{k}}] + h_{\mathrm{A}}(\tilde{S}_{\vec{k}})S_{\mathrm{fail}}. \tag{E.9}$$

Note that condition (ii) of the theorem tells us that the dimension of the typical subspace is at most $2^{n(\widehat{\mathrm{H}}(\mathsf{Src})+\varepsilon)}$. For any $R > \widehat{\mathrm{H}}(\mathsf{Src})$, we can therefore find an ε such that we achieve a compression of rate R.

To decompress, we will do nothing and simply output

$$\check{S}_{\vec{k}} := \hat{S}_{\vec{k}}, \tag{E.10}$$

and so all that remains is to show that $\hat{S}_{\vec{k}}$ is in fact close to the original state $\tilde{S}_{\vec{k}}$. Suppose for simplicity that the maximum is attained when computing the distance, and let $\mathbf{e}$ denote the optimal measurement. That is

$$\mathcal{D}(\hat{S}_{\vec{k}}, \tilde{S}_{\vec{k}}) = \sup_{\mathbf{f}} \mathcal{C}(\mathbf{f}(\hat{S}_{\vec{k}}), \mathbf{f}(\tilde{S}_{\vec{k}})) = \mathcal{C}(\mathbf{e}(\hat{S}_{\vec{k}}), \mathbf{e}(\tilde{S}_{\vec{k}})), \tag{E.11}$$

We then have

$$\mathcal{D}(\check{S}_{\vec{k}}, \tilde{S}_{\vec{k}}) = \mathcal{C}(\mathbf{e}(\hat{S}_{\vec{k}}), \mathbf{e}(\tilde{S}_{\vec{k}})) \tag{E.12}$$

$$\leq h_T(\tilde{S}_{\vec{k}})\mathcal{C}(\mathbf{e}(T[\tilde{S}_{\vec{k}}]), \mathbf{e}(\tilde{S}_{\vec{k}})) + h_A(\tilde{S}_{\vec{k}})\mathcal{C}(\mathbf{e}(S_{\mathrm{fail}}), \mathbf{e}(\tilde{S}_{\vec{k}})) \tag{E.13}$$

$$\leq h_T(\tilde{S}_{\vec{k}})\mathcal{D}(T[\tilde{S}_{\vec{k}}], \tilde{S}_{\vec{k}}) + h_A(\tilde{S}_{\vec{k}})\mathcal{D}(S_{\mathrm{fail}}, \tilde{S}_{\vec{k}}) \tag{E.14}$$

$$\leq c h_A(\tilde{S}_{\vec{k}})^{\varepsilon} + h_A(\tilde{S}_{\vec{k}}), \tag{E.15}$$

$$\leq (c+1)h_A(\tilde{S}_{\vec{k}})^{\varepsilon} \tag{E.16}$$

where the first inequality follows from the properties of the classical trace distance and the linearity of effects, the second from the definition of distance, and the third from the weak disturbance property of a pseudo-projective measurement, where $c \geqslant 0$ and $\varepsilon \in (0, 1]$ are constants given by a particular theory. We then note that

$$\sum_{\vec{k}} q_{\vec{k}} \mathcal{D}(\tilde{S}_{\vec{k}}, \breve{S}_{\vec{k}}) \leq \left(\sum_{\vec{k}} q_{\vec{k}} \mathcal{D}(\tilde{S}_{\vec{k}}, \breve{S}_{\vec{k}})^{\frac{1}{\varepsilon}} \right)^{\varepsilon} \tag{E.17}$$

$$\leq \left(\sum_{\vec{k}} q_{\vec{k}} (c+1)^{\frac{1}{\varepsilon}} h_A(\tilde{S}_{\vec{k}}) \right)^{\varepsilon} \tag{E.18}$$

$$= \left((c+1)^{\frac{1}{\varepsilon}} h_A(\mathsf{Src}^{\otimes n}) \right)^{\varepsilon} \tag{E.19}$$

$$\leq (c+1)\delta^{\varepsilon} \tag{E.20}$$

The inequality in the last line follows from the typical subspace theorem. As δ can be chosen to be arbitrarily small, this concludes our proof. □

References

[1] Barnum H, Barrett J, Leifer M and Wilce A 2007 Generalized no-broadcasting theorem *Phys. Rev. Lett.* **99** 240501

[2] Barnum H, Dahlsten O C O, Leifer M and Toner B 2008 Nonclassicality without entanglement enables bit commitment *Proc. IEEE ITW* pp. 386–90

[3] Barrett J 2007 Information processing in generalized probabilistic theories *Phys. Rev.* A **75** 032304

[4] Brandao F G S L and Plenio M B 2009 Generalization of quantum Stein's lemma. arxiv:0904.0281

[5] Brassard G 2005 Is information the key? *Nat. Phys.* **1** 2–4

[6] Buhrman H, Christandl M, Unger F, Wehner S and Winter A 2006 Implications of superstrong nonlocality for cryptography *Proc. R. Soc.* A **462** 1919–32

[7] Fuchs C A 2002 Quantum mechanics as quantum information (and only a little more). arxiv:quant-ph/0205039

[8] Cerf N J and Adami C 1997 Negative entropy and information in quantum mechanics *Phys. Rev. Lett.* **79** 5194

[9] Clauser J, Horne M, Shimony A and Holt R 1969 Proposed experiment to test local hidden-variable theories *Phys. Rev. Lett.* **23** 880–4

[10] Clifton R, Bub J and Halvorson H 2003 Characterizing quantum theory in terms of information-theoretic constraints *Found. Phys.* **33** 1561–91

[11] D'Ariano G M 2008 Probabilistic theories: what is special about quantum mechanics? arXiv:0807.4383

[12] Datta A, Shaji A and Caves C M 2008 Quantum discord and the power of one qubit *Phys. Rev. Lett.* **100** 050502

[13] Doherty A C and Wehner S 2007 unpublished

[14] Fannes M 1973 A continuity property of the entropy density for spin lattice systems *Commun. Math. Phys.* **31** 291–4

[15] Hänggi E, Renner R and Wolf S 2009 The impossibility of non-signaling privacy amplification. arXiv:0906.4760

[16] Hardy L 2001 Quantum theory from five reasonable axioms. arXiv:quant-ph/0101012

[17] Hayashi M 2006 *Quantum Information: an Introduction* (Berlin: Springer)

[18] Hiai F and Petz D 1991 The proper formula for the relative entropy and its asymptotics in quantum probability *Commun. Math. Phys.* **143** 99–114

[19] Horn R A and Johnson C R 1985 *Matrix Analysis* (Cambridge: Cambridge University Press)

[20] Horodecki M, Oppenheim J and Winter A 2005 Quantum information can be negative *Nature* **436** 673–6

[21] Horodecki M, Oppenheim J and Winter A 2007 Quantum state merging and negative information *Commun. Math. Phys.* **269** 107

[22] Matthews W, Wehner S and Winter A 2009 Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding *Commun. Math. Phys.* **291** 813–43

[23] Nayak A 1999 Optimal lower bounds for quantum automata and random access codes *Proc. 40th IEEE FOCS* pp 369–76 (arXiv:quant-ph/9904093)

[24] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)

[25] Ogawa T and Nagaoka H 2000 Strong converse and Stein's lemma in quantum hypothesis testing *IEEE Trans. Inf. Theory* **46** 2428

[26] Pawlowski M, Paterek T, Kaszlikowski D, Scarani V, Winter A and Zukowski M 2009 A new physical principle: information causality arXiv:0905.2292

[27] Popescu S and Rohrlich D 1994 Quantum nonlocality as an axiom *Found. Phys.* **24** 379–85

[28] Popescu S and Rohrlich D 1996 Nonlocality as an axiom for quantum theory *The Dilemma of Einstein, Podolsky and Rosen, 60 Years Later: Int. Symp. in Honour of Nathan Rosen*

[29] Popescu S and Rohrlich D 1997 Causality and nonlocality as axioms for quantum mechanics *Proc. Symp. of Causality and Locality in Modern Physics and Astronomy: Open Questions and Possible Solutions*

[30] Schumacher B 1995 Quantum coding *Phys. Rev.* A **51** 2738

[31] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379–423, 623–56

[32] Short A J and Barrett J 2009 Strong nonlocality: a trade-off between states and measurements arXiv:0909.2601

[33] Short A J, Gisin N and Popescu S 2006 The physics of no-bit-commitment: generalized quantum non-locality versus oblivious transfer *Quantum Inf. Process.* **5** 1573

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

[35] Ver Steeg G and Wehner S 2009 Relaxed uncertainty relations and information processing *Quantum Inf. Comput.* **9** 801

[36] Tsirelson B 1980 Quantum generalizations of Bell's inequality *Lett. Math. Phys.* **4** 93–100

[37] van Dam W 2005 Impossible consequences of superstrong nonlocality arXiv:quant-ph/0501159

[38] Wehner S, Christandl M and Doherty A C 2008 A lower bound on the dimension of a quantum system given measured data *Phys. Rev.* A **78** 062112

[39] Wolf S and Wullschleger J 2005 Bit commitment from weak non-locality arXiv:quant-ph/0508233

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

[41] Kimura G, Nuida K and Imai H 2009 Distinguishability measures and entropies for general probabilistic theories arXiv:0910.0994

[42] Barrett J, Linden N, Massar S, Pironio S, Popescu S and Roberts D 2005 Nonlocal correlations as an information-theoretic resource *Phys. Rev.* A **71** 022101

[43] Winter A 1999 Coding theorem and strong converse for quantum channels *IEEE Trans. Inf. Theory* **45** 2481