On the Supposed Conceptual Inadequacy of the Shannon Information

C.G. Timpson
The Queen's College, Oxford, OX1 4AW, UK[*]

(Dated: 30 December 2001)

In Phys. Rev. A **63**, 022113 (2001), Brukner and Zeilinger claim that the Shannon information is not well defined as a measure of information in quantum mechanics, adducing arguments that seek to show that it is inextricably tied to classical notions of measurement. It is shown here that these arguments do not succeed: the Shannon information does not have problematic ties to classical concepts. In a further argument, Brukner and Zeilinger compare the Shannon information unfavourably to their preferred information measure, $I(\vec{p})$, with regard to the definition of a notion of 'total information content'. This argument is found unconvincing and the relationship between individual measures of information and notions of 'total information content' investigated.

I. INTRODUCTION

What role the concept of information might have to play in the foundations of quantum mechanics is a question that has recently excited renewed interest [1–3]. Zeilinger, for example, has put forward an information-theoretic principle which he suggests might serve as a foundational principle for quantum mechanics [4]. As a part of this project, Brukner and Zeilinger [5] have criticised the Shannon measure of information [6], the quantity fundamental to the discussion of information in both classical and quantum information theory. They claim that the Shannon information is not appropriate as a measure of information in the quantum context and have proposed in its stead their own preferred quantity and a notion of 'total information content' associated with it, which latter is supposed to supplant the von Neumann entropy [8–10].

Their main aim in [5] is to establish that the Shannon information is intimately tied to classical notions, in particular, to the preconceptions of classical measurement, and that in consequence it cannot serve as a measure of information in the quantum context. They seek to establish this in two ways. First, by arguing that the Shannon measure only makes sense when we can take there to be a pre-existing sequence of bit values in a message we are decoding, which is not the case in general for measurements on quantum systems (consider measurements on qubits in a basis different from their eigenbasis); and second, by suggesting that Shannon's famous third postulate, the postulate that secures the uniqueness of the form of the Shannon information measure [6] and has been seen by many as a necessary axiom for a measure of information, is motivated by classical preconceptions and does not apply in general in quantum mechanics where we must consider non-commuting observables.

These two arguments do not succeed in showing that the Shannon information is 'intimately tied to the notion of systems carrying properties prior to and independent of observation'[10], however. The first is based on too narrow a conception of the meaning of the Shannon information and the second, primarily, on a misreading of what is known as the 'grouping axiom'. We shall see that the Shannon information is perfectly well defined and appropriate as a measure of information in the quantum context as well as in the classical. We will begin by reviewing some of the different ways in which the Shannon information may be understood (Section II), before examining this pair of arguments and seeing where they go wrong (Section III).

Brukner and Zeilinger have a further argument against the Shannon information (Section IV). They suggest it is inadequate because it cannot be used to define an acceptable notion of 'total information content'. Equally, they insist, the von Neumann entropy cannot be a measure of information content for a quantum system because it has no general relation to information gain from the measurements that we might perform on a system, save in the case of measurement in the basis in which the density matrix is diagonal. By contrast, for a particular set of measurements, their preferred information measure sums to a unitarily invariant quantity that they interpret as 'information content', this being one of their primary reasons for adopting this specific measure. This property will be seen to have a simple geometric explanation in the Hilbert-Schmidt representation of density operators however, rather than being of any great information theoretic significance; and this final argument found unpersuasive, as the proposed constraint on any information measure regarding the definition of 'total information content' seems unreasonable. Part of the problem is that information content, total or otherwise, is not a univocal concept and we need to be careful to specify precisely what we might mean by it in any given context.

II. INTERPRETATION OF THE SHANNON INFORMATION

The technical concept of information relevant to our discussion, the Shannon information, finds its home in the context of communication theory. We are concerned with a notion of *quantity* of information; and the notion of quantity of information is cashed out in terms of the

[*]Electronic address: `christopher.timpson@queens.ox.ac.uk`

resources required to transmit messages (which is, note, a very limited sense of quantity). We shall highlight two main ways in which the Shannon information may be understood, the first of which rests explicitly on Shannon's 1948 noiseless coding theorem [6].

A. The communication channel

It is instructive to begin with a quotation of Shannon's:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently these messages have *meaning*...These semantic aspects of communication are irrelevant to the engineering problem. [6, p.31]

The communication system consists of an information source, a transmitter or encoder, a (possibly noisy) channel, and a receiver (decoder). It must be able to deal with *any* possible message produced (a string of symbols selected in the source, or some varying waveform), hence it is quite irrelevant whether what is actually transmitted has any meaning or not.

It is crucial to realise that 'information' in Shannon's theory is not associated with individual messages, *but rather characterises the source of the messages.* The point of characterising the source is to discover what capacity is required in a communications channel to transmit all the messages the source produces; and it is for this that the concept of the Shannon information is introduced. The idea is that the statistical nature of a source can be used to reduce the capacity of channel required to transmit the messages it produces (we shall restrict ourselves to the case of discrete messages for simplicity).

Consider an ensemble X of letters $\{x_1, x_2, \ldots, x_n\}$ occurring with probabilities p_i. This ensemble is our source, from which messages of N letters are drawn. We are concerned with messages of very large N. For such messages, we know that typical sequences of letters will contain Np_i of letter x_i, Np_j of x_j and so on. The number of distinct typical sequences of letters is then given by

$$\frac{N!}{Np_1!Np_2!\ldots Np_n!}$$

and using Stirling's approximation, this becomes $2^{NH(X)}$ where

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

is the Shannon information (logarithms are to base 2 to fix the units of information as binary bits).

Now as $N \to \infty$, the probability of an atypical sequence appearing becomes negligible and we are left with only $2^{NH(X)}$ equiprobable typical sequences which need ever be considered as possible messages. We can thus replace each typical sequence with a binary code number of $NH(X)$ bits and send that to the receiver rather than the original message of N letters ($N \log n$ bits).

The message has been compressed from N letters to $NH(X)$ bits ($\leq N \log n$ bits). Shannon's noiseless coding theorem, of which this is a rough sketch, states that this represents the optimal compression [6]. The Shannon information is, then, appropriately called a measure of information because it represents the maximum amount that messages consisting of letters drawn from an ensemble X can be compressed.

One may also make the derivative statement that the information *per letter* in a message is $H(X)$ bits, which is equal to the information of the source. But 'derivative' is an important qualification: we can only consider a letter x_i drawn from an ensemble X to have associated with it the information $H(X)$ if we consider it to be a member of a typical sequence of N letters, where N is large, drawn from the source.

Note also that we must strenuously resist any temptation to conclude that because the Shannon information tells us the maximum amount a message drawn from an ensemble can be compressed, that it therefore tells us the irreducible meaning content of the message, specified in bits, which somehow possess their own intrinsic meaning. This idea rests on a failure to distinguish between a code, which has no concern with meaning, and a language, which does (cf. [11] Chpt.5).

B. Information and Uncertainty

Another way of thinking about the Shannon information is as a measure of the amount of information that we *expect* to gain on performing a probabilistic experiment. The Shannon measure is a measure of the uncertainty of a probability distribution as well as serving as a measure of information. A measure of uncertainty is a quantitative measure of the lack of concentration of a probability distribution; this is called an uncertainty because it measures our uncertainty about what the outcome of an experiment completely described by the probability distribution in question will be. Uffink provides an axiomatic characterisation of measures of uncertainty, deriving a general class of measures of which the Shannon information is one [12, 13].

Imagine a truly random probabilistic experiment described by a probability distribution $\vec{p} = \{p_1, \ldots, p_n\}$. The intuitive link between uncertainty and information is that the greater the uncertainty of this distribution, the more we stand to gain from learning the outcome of the experiment. In the case of the Shannon information, this notion of how much we gain can be made more precise.

Some care is required when we ask 'how much do we know about the outcome?' for a probabilistic experiment. In a certain sense, the shape of the probability distribution might provide no information about what an individual outcome will actually be, as any of the outcomes assigned non-zero probability can occur. However, we can use the probability distribution to put a *value* on any given outcome. If it is a likely one, then it will be no surprise if it occurs so of little value; if an unlikely one, it is a surprise, hence of higher value. A nice measure for the value of the occurrence of outcome i is $-\log p_i$, a decreasing function of the probability of the outcome. We may call this the 'surprise' information associated with outcome i; it measures the value of what we learn from the experiment given that we know the probability distribution for the outcomes.

If the information that we would gain if outcome i were to occur is $-\log p_i$, then before the experiment, the amount of information we expect to gain is given by the expectation value of the 'surprise' information, $\sum_i p_i(-\log p_i)$; and this, of course, is just the Shannon information H of the probability distribution $\vec{p}$. Hence the Shannon information tells us our expected information gain.

More generally, a crude sketch of how the relationship between uncertainty and expected information gain might be cashed out for the whole class of measures of uncertainty may be given as follows. What we know given the probability distribution for an experiment is that if the experiment is repeated very many times then the sequence of outcomes we attain will be one of the typical sequences. How much we learn from actually performing the experiments and acquiring one of those sequences, then, will depend on the number of typical sequences; the more there are, the more we stand to gain. Thus for a quantitative measure of how much information we gain from the sequence of experiments we could just count the number of typical sequences (which would give us NH, the Shannon information of the sequence), or we could choose any suitably behaved function (e.g. continuous, invariant under relabelling of the outcome probabilities) that increases as the number of typical sequences increases. This property will follow from Schur concavity, which is the key requirement on Uffink's general class of uncertainty measures $U_r(\vec{p})$ (for details of the property of Schur concavity, see Refs. 12, 14 and Section IV A below). $NU_r(\vec{p})$ then, can be understood as a measure of the amount of information we gain from a long series N of experiments; we get the average or expected information per measurement by dividing by N, but note that this quantity only makes sense if we consider the individual measurement as part of a long sequence of measurements.

A precisely similar story can be told for a measure of 'how much we know' given a probability distribution. This will be the inverse of an uncertainty: we want a measure of the concentration of a probability distribution; the more concentrated, the more we know about what the outcome will be (which just means, the better we can predict the outcome). A function that decreases as the number of typical sequences increases (Schur convexity) will give our quantitative measure of how much we know about what the outcome of a long run of experiments will be: the more typical sequences there are the less we know. We are again talking of a long run of experiments, so 'how much we know' for a single experiment will only make sense as an average value, when the single experiment is considered as a member of a long sequence. So note again that to say we have a certain amount of information (knowledge) about what the outcome of an experiment will be is not to claim that we have partial knowledge of some predetermined fact about the outcome of an experiment.

C. The minimum number of questions needed to specify a sequence

The final common interpretation of the Shannon information is as the minimum average number of binary questions needed to specify a sequence drawn from an ensemble [12, 15], although this appears not to provide an interpretation of the Shannon information actually independent of the previous two.

Imagine that a long sequence N of letters is drawn from the ensemble X, or that N independent experiments whose possible outcomes have probabilities p_i are performed, but the list of outcomes is kept from us. Our task is to determine what the sequence is by asking questions to which the guardian of the sequence can only answer 'yes' or 'no'; and we choose to do so in such a manner as to minimize the average number of questions needed. We need to be concerned with the *average* number to rule out lucky guesses identifying the sequence.

If we are trying to minimize the average number of questions, it is evident that the best questioning strategy will be one that attempts to rule out half the possibilities with each question, for then whatever the answer turns out to be, we still get the maximum value from each question. Given the probability distribution, we may attempt to implement this strategy by dividing the possible outcomes of each individual experiment into classes of equal probability, and then asking whether or not the outcome lies in one of these classes. We then try and repeat this process, dividing the remaining set of possible outcomes into two sets of equal probabilities, and so on. It is in general not possible to proceed in this manner, dividing a finite set of possible outcomes into two sets of equal probabilities, and it can be shown that in consequence the average number of questions required if we ask about each individual experiment in isolation is greater than or equal to $H(X)$. However, if we consider the N repeated experiments, where N tends to infinity, and consider asking joint questions about what the outcomes of the independent experiments were, we can always divide the classes of possibilities of (joint) outcomes in the required way. Now we already know that for large N, there are $2^{NH(X)}$ typical sequences, so given that we can strike out half the possible sequences with each question, the minimum average number of questions needed to identify the sequence is $NH(X)$. (These last results are again essentially the noiseless coding theorem.)

It is not immediately obvious, however, why the minimum average number of questions needed to specify a sequence should be related to the notion of information. (Again, the tendency to think of bits and binary questions as irreducible meaning elements is to be resisted.) It seems, in fact that this is either just another way of talking about the maximum amount that messages drawn from a given ensemble can be compressed, in which case we are back to the interpretation of the Shannon information in terms of the noiseless coding theorem, or it is providing a particular way of characterising how much we stand to gain from learning a typical sequence, and we return to an interpretation in terms of our expected information gain.

III. TWO ARGUMENTS AGAINST THE SHANNON INFORMATION

With this preamble behind us, we may turn to the first of the arguments against the Shannon information.

A. Are pre-existing bit-values required?

Since the quantity $-\sum_i p_i \log p_i$ is meaningful for any (discrete) probability distribution (and can be generalised for continuous distributions), Brukner and Zeilinger's argument must be that when we have probabilities arising from measurements on quantum systems, $-\sum_i p_i \log p_i$ does not correspond to a concept of *information*. Their argument concerns measurements on systems that are all prepared in a given state $|\psi\rangle$, where $|\psi\rangle$ may not be an eigenstate of the observable we are measuring. The probability distribution $\vec{p} = \{p_1, \dots, p_n\}$ for measurement outcomes will be given by $p_i = \mathrm{Tr}(|\psi\rangle\langle\psi|P_i)$, where P_i are the operators corresponding to different measurement outcomes (projection operators in the spectral decomposition of the observable, for projective measurements).

Brukner and Zeilinger suggest that the Shannon information has no meaning in the quantum case, because the concept lacks an 'operational definition' in terms of the number of binary questions needed to specify an actual concrete sequence of outcomes. In general in a sequence of measurements on quantum systems, we cannot consider there to be a pre-existing sequence of possessed values, at least if we accept the orthodox eigenvalue-eigenstate link for the ascription of definite values [16][1], and this rules out, they insist, interpreting the Shannon measure as an amount of information:

> The nonexistence of well-defined bit values prior to and independent of observation suggests that the Shannon measure, as defined by the number of binary questions needed to determine the particular *observed* sequence 0's and 1's, becomes problematic and even untenable in defining our uncertainty as given *before* the measurements are performed. [5, p.1]

> ...No definite outcomes exists [sic] before measurements are performed and therefore the number of different possible sequences of outcomes does not characterize our uncertainty about the individual system before measurements are performed. [*ibid.* p.3]

These two statements should immediately worry us, however. Recall the key points of the interpretation of the Shannon information: given a long message (a long run of experiments), we *know* that it will be one of the typical sequences that is instantiated. Given $\vec{p}$, we can say what the typical sequences will be, how many there are, and hence the number of bits $(NH(X))$ needed to specify them, independent of whether or not there is a pre-existing sequence of bit values. It is irrelevant whether there already is some concrete sequence of bits or not; all possible sequences that will be produced will require the same number of bits to specify them as any sequence produced will always be one of the typical sequences. It clearly makes no difference to this whether the probability distribution is given classically or comes from the trace rule. Also, the number of different possible sequences does indeed tell us about our uncertainty before measurement: what we know is that one of the typical sequences will be instantiated, what we are ignorant of is which one it will be, and we can put a measure on how ignorant we are simply by counting the number of different possibilities. Brukner and Zeilinger's attempted distinction between uncertainty before and after measurement is not to the point, the uncertainty is a function of the probability distribution and this is perfectly well defined before measurement[2].

[1] In a footnote, Brukner and Zeilinger suggest that the Kochen-Specker theorem in particular raises problems for the operational definition of the Shannon information. It is not clear, however, why the impossibility of assigning *context independent* yes/no answers to questions asked of the system should be a problem if we are considering an *operational* definition. Presumably such a definition would include a concrete specification of the experimental situation, i.e. refer to the context, and then we are not concerned with assigning a value to an *operator* but to the outcome of a specified experimental procedure, and this can be done perfectly consistently, if we so wish. The de-Broglie Bohm theory, of course, provides a concrete example [17].

[2] We may need to enter at this point the important note that the Shannon information is not supposed to describe our *general* uncertainty when we know the *state*, this is a job for a measure of mixedness such as the von Neumann entropy, see below.

Brukner and Zeilinger have assumed that it is a necessary and sufficient condition to understand H as a measure of information that there exists some concrete string of N values, for then and only then can we talk of the minimum number of binary questions needed to specify the string. But as we have now seen, it is *not* a necessary condition that there exist such a sequence of outcomes.

We are not in any case forced to assume that H is about the number of questions needed to specify a sequence in order to understand it as a measure of information; we also have the interpretations in terms of the maximum amount a message drawn from an ensemble described by the probability distribution $\vec{p}$ can be compressed, and as the expected information gain on measurement. (And as we have seen, one of these two interpretations must in fact be prior.) Furthermore, the absence of a pre-existing string need not even be a problem for the minimum average questions interpretation - we can ask about the minimum average number of questions that *would* be required if we *were* to have a sequence drawn from the ensemble. So again, the pre-existence of a definite string of values is not a necessary condition.

It is not a sufficient condition either, because, faced with a string of N definite outcomes, in order to interpret NH as the minimum *average* number of questions needed to specify the sequence, we need to know that we in fact have a typical sequence, that is, we need to imagine an ensemble of such typical sequences and furthermore, to assume that the relative frequencies of each of the outcomes in our actual string is representative of the probabilities of each of the outcomes in the notional ensemble from which the sequence is drawn. If we do not make this assumption, then the minimum number of questions needed to specify the state of the sequence must be N - we cannot imagine that the statistical nature of the source from which the sequence is notionally drawn allows us to compress the message. So even in the classical case, the concrete sequence on its own is not enough and we need to consider an ensemble, either of typical sequences or an ensemble from which the concrete sequence is drawn. In this respect the quantum and classical cases are completely on a par. The same assumption needs to be made in both cases, namely, that the probability distribution $\vec{p}$, either known in advance, or derived from observed relative frequencies, correctly describes the probabilities of the different possible outcomes. The fact that no determinate sequence of outcomes exists before measurement does not pose any problems for the Shannon information in the quantum context.

Reiterating their requirements for a satisfactory notion of information, Brukner and Zeilinger say:

> We require that the information gain be directly based on the observed probabilities, (and not, for example, on the precise sequence of individual outcomes observed on which Shannon's measure of information is based). [10]

But as we have seen, it is false that the Shannon measure must be based on a precise sequence of outcomes (this is not a necessary condition) and the Shannon measure already *is* and *must be* based on the observed probabilities (a sequence of individual outcomes on its own is not sufficient).

There is, however, a difference between the quantum and classical cases that Brukner and Zeilinger may be attempting to capture. Suppose we have a sequence of N qubits that has actually been used to encode some information, that is, the sequence of qubits is a channel to which we have connected a classical information source. For simplicity, imagine that we have coded in orthogonal states. Then the state of the sequence of qubits will be a product of $|0\rangle$'s and $|1\rangle$'s and for measurements in the encoding basis, the sequence will have a Shannon information equal to $NH(A)$ where $H(A)$ is the information of the classical source. If we do not measure in the encoding basis, however, the sequence of 0's and 1's we get as our outcomes will differ from the values originally encoded and the Shannon information of the resulting sequence will be greater than that of the original[3]. We have introduced some 'noise' by measuring in the wrong basis. The way we describe this sort of situation, though [19], is to use the Shannon mutual information $H(A:B) = H(A) - H(A|B)$, where B denotes the outcome of measurement of the chosen observable (outcomes b_i with probabilities $p(b_i)$) and the 'conditional entropy' $H(A|B) = \sum_{i=1}^{n} p(b_i)H(p(a_1|b_i), \ldots, p(a_m|b_i))$, characterises the noise we have introduced by measuring in the wrong basis. $H(B)$ is the information (per letter) of the sequence that we are left with after measurement, $H(A:B)$ tells us the amount of information that we have actually managed to transmit down our channel, i.e. the amount (per letter) that can be decoded when we measure in the wrong basis.

B. The grouping axiom

The first argument has not revealed any difficulties for the Shannon information in the quantum context, so let us now turn to the second.

In his original paper, Shannon put forward three properties as reasonable requirements on a measure of uncertainty and showed that the only function satisfying these requirements has the form $H = -K\sum_i p_i \log p_i$.

---

[3] We may think of our initial sequence of qubits as forming an ensemble described by the density operator $\rho = p_1|0\rangle\langle 0| + p_2|1\rangle\langle 1|$, where p_1, p_2 are the probabilities for 0 and 1 in our original classical information source. Any (projective) measurement that does not project onto the eigenbasis of ρ will result in a post-measurement ensemble that is more mixed than ρ (see e.g. Refs. 14, 18 and below) and hence will have a greater uncertainty, thus a greater Shannon information, or any other measure of information gain.

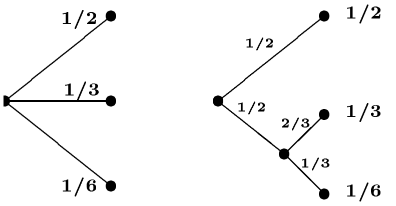

FIG. 1: The probability distribution $\vec{p} = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\}$ can be considered as given for the three outcomes directly, or we could consider first a choice of two equiprobable events, followed by a second choice of two events with probabilities $\frac{2}{3}, \frac{1}{3}$, conditional on the second, say, of the first two events occurring, a 'decomposition' of a single choice into two successive choices, the latter of which will only be made half the time. Shannon's third requirement says that the uncertainty in $\vec{p}$ will be given by $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$: the uncertainty of the overall choice is equal to the uncertainty of the first stage of the choice, plus the uncertainty of the second choice weighted by its probability of occurrence.

(In contrast to some later writers, however, notably Jaynes [20], he set little store by this derivation, seeing the justification of his measure as lying rather in its implications [6]. Save the noiseless coding theorem, the most significant of the implications that Shannon goes on to draw are, as has been pointed out by Uffink, consequences of the property of Schur concavity and hence shared by the general class of measures of uncertainty derived in [12].)

The first two requirements are that H should be continuous in the p_i and that for equiprobable events ($p_i = 1/n$), H should be a monotonic increasing function of n. The third requirement is the strongest and the most important in the uniqueness proof. It states that if a choice is broken down into two successive choices, the original H should be a weighted sum of the individual values of H. The meaning of this rather non-intuitive constraint is usually demonstrated with an example (see Fig. 1). A precise statement of Shannon's third requirement (one that includes also the second requirement as a special case) is due to Faddeev [21] and is often known as the Faddeev grouping axiom:

Grouping Axiom 1 (Faddeev) *For every $n \geq 2$*

$$H(p_1, p_2, \ldots, p_{n-1}, q_1, q_2) = H(p_1, \ldots, p_{n-1}, p_n) + p_n H(\frac{q_1}{p_n}, \frac{q_2}{p_n})$$

where $p_n = q_1 + q_2$.

The form of the Shannon information follows uniquely from requiring $H(p, 1-p)$ to be continuous for $0 \leq p \leq 1$ and positive for at least one value of p, permutation invariance of H with respect to relabelling of the p_i, and the grouping axiom.

'Grouping axiom' is an appropriate name. As it is standardly understood (see e.g. [12, 15, 20]), we consider that instead of giving the probabilities $p_1, \ldots, p_n$ of the outcomes $x_1, \ldots, x_n$ of a probabilistic experiment directly, we may imagine grouping the outcomes into composite events whose probabilites will be given by the sum of the probabilities of their respective component events, and then specifying the probabilities of the outcome events conditional on the occurrence of the composite events to which they belong; this way of specifying the probabilistic experiment being precisely equivalent to the first. So we might group the first k events together into an event A, which would have a probability $p(A) = \sum_{i=1}^{k} p_i$, and the remaining $n-k$ into an event B of probability $p(B) = \sum_{i=k+1}^{n} p_i$; and then give the conditional probabilities of the events $x_1, \ldots, x_k$ conditional on composite event A occurring, $(p_1/p(A)), \ldots, (p_k/p(A))$, and similarly the conditional probabilities for the events $x_{k+1}, \ldots, x_n$ conditional on event B. The grouping axiom then concerns how the uncertainty measures should be related for these different descriptions of the same probabilistic experiment. It says that our uncertainty about which event will occur should be equal to our uncertainty about which group it will belong to plus the expected value of the uncertainty that would remain if we were to know which group it belonged to (this expected value being the weighted sum of the uncertainties of the conditional distributions, with weights given by the probability of the outcome lying within a given group).

So in particular, let us imagine an experiment with $n+1$ outcomes which we label $a_1, a_2, \ldots, a_{n-1}, b_1, b_2$, having probabilities $p_1, \ldots, p_{n-1}, q_1, q_2$ respectively. We can define an event $a_n = b_1 \cup b_2, b_1 \cap b_2 = \emptyset$, which would have probability $p_n = q_1 + q_2$ and the probabilities for b_1 and b_2 conditional on a_n occurring will then be $\frac{q_1}{p_n}, \frac{q_2}{p_n}$ respectively. Grouping Axiom 1 says that the uncertainty in the occurrence of events $a_1, a_2, \ldots, a_{n-1}, b_1, b_2$ is equal to the uncertainty for the occurrence of events $a_1, \ldots, a_n$ plus the uncertainty for the occurrence of b_1, b_2 conditional on a_n occurring, weighted by the probability that a_n should occur.

Brukner and Zeilinger suggest that the grouping axiom, however, embodies certain classical presumptions that do not apply in quantum mechanics. This entails that the axiomatic derivation of the form of the Shannon measure does not go through and that the Shannon information ceases to be a measure of uncertainty in the quantum context. The argument turns on their interpretation of the grouping axiom, which differs from the standard interpretation in that it refers to joint experiments.

1. *Brukner and Zeilinger's interpretation*

If we take an experiment, A, with outcomes $a_1, \ldots, a_n$ and probabilities $(p(a_1), \ldots, p(a_n)) = (p_1, \ldots, p_n)$ and an experiment, B, with outcomes b_1, b_2, then for the joint experiment $A \wedge B$, the event a_n is the union of the two disjoint events $a_n \wedge b_1$ and $a_n \wedge b_2$. Let us assign to these two events the probabilities q_1 and q_2 respectively. Then $p(a_n) = p(a_n \wedge b_1) + p(a_n \wedge b_2) = q_1 + q_2 = p_n$. On this interpretation, the left hand side of Grouping Axiom 1 is to be understood as denoting the uncertainty in the experiment with outcomes $a_1, a_2, \ldots, a_{n-1}, a_n \wedge b_1, a_n \wedge b_2$.

If a_n occurs, the conditional probabilities for b_1, b_2 will be $p(a_n \wedge b_1)/p(a_n) = q_1/p_n, p(a_n \wedge b_2)/p(a_n) = q_2/p_n$ respectively, and so $H(\frac{q_1}{p_n}, \frac{q_2}{p_n})$ is the uncertainty in the value of B given that a_n occurs.

The grouping axiom can now be rewritten as:

Grouping Axiom 2 (Brukner and Zeilinger)

$$\begin{aligned} H\left(p(a_1), p(a_2), \ldots, p(a_{n-1}), p(a_n \wedge b_1), p(a_n \wedge b_2)\right) \\ = H\left(p(a_1), p(a_2), \ldots, p(a_n)\right) \\ + p(a_n) H\left(p(b_1|a_n), p(b_2|a_n)\right). \end{aligned}$$

Generalizing to the case in which we have m outcomes for experiment B and distinguish B values for all n A outcomes, so that we have mn outcomes $a_i \wedge b_j$, the grouping axiom becomes:

Grouping Axiom 3 (Brukner and Zeilinger)

$$H(A \wedge B) = H(A) + H(B|A)$$

From the point of view of Shannon's original presentation, this expression appears as a theorem rather than an axiom, being a consequence of the logarithmic form of the Shannon information and the definition of the conditional entropy.

2. *The inapplicability argument*

The classical assumptions made explicit, Brukner and Zeilinger suggest, in Grouping Axioms 2 and 3 are that attributes corresponding to all possible measurements can be assigned to a system simultaneously (in this case, a_i, b_j and $a_i \wedge b_j$); and that measurements can be made ideally non-disturbing. Grouping Axiom 3, for example, is supposed to express the fact that classically, the information we expect to gain from a joint experiment $A \wedge B$, is the same as the information we expect to gain from first performing A, then performing B (where the uncertainty in B is updated conditional on the A outcome, but our ability to predict B outcomes is not degraded by the A measurement).

Their inapplicability argument is simply that as the grouping axiom requires us to consider joint experiments, the uniqueness proof for the Shannon information will fail in the quantum context, because we can consider measurements of non-commuting observables and the joint probabilities on the left hand side of Grouping Axiom 2 will not be defined for such observables; thus the grouping axiom will fail to hold. Furthermore, the grouping axiom shows that the Shannon information embodies classical assumptions, so the Shannon measure will not be justified as a measure of uncertainty because these assumptions do not hold in the quantum case. The result is that

> ...only for the special case of commuting, i.e., simultaneously definite observables, is the Shannon measure of information applicable *and the use of the Shannon information justified to define the uncertainty given before quantum measurements are performed.* ([5] p.4, my emphasis)

This argument is problematic, however. Let us begin with the obvious point that a failure of the argument for uniqueness does not automatically rule out the Shannon information as a measure of uncertainty. In fact, the Shannon information can be seen as one of a general class of measures of uncertainty, characterised by a set of axioms in which the grouping axiom does not appear [12], hence the grouping axiom is not necessary for the interpretation of the Shannon information as a measure of uncertainty. (Uffink in fact has previously argued that the grouping axiom is not a natural constraint on a measure of information and should not be imposed as a necessary constraint, even in the classical case ([12], §1.6.3).) So from the fact that on the Brukner/Zeilinger reading, the grouping axiom seems to embody some classical assumptions that do not hold in the quantum case, it does not follow that the concept of the Shannon information as a measure of uncertainty involves those classical assumptions.

Furthermore, Brukner and Zeilinger's grouping axiom is not in fact equivalent to the standard form and the standard form is equally applicable in both the classical and quantum cases. Thus the Shannon information has not been shown to involve classical assumptions and the standard axiomatic derivation can indeed go through in the quantum context. The probabilities appearing in Grouping Axiom 1 are well defined in both the classical and quantum cases.

In Brukner and Zeilinger's notation, Grouping Axiom 1 would be written as

$$\begin{aligned} & H(p(a_1), p(a_2), \ldots, p(a_{n-1}), p(b_1), p(b_2)) \\ = \; & H(p(a_1), \ldots, p(a_{n-1}), p(b_1 \vee b_2)) \\ & + p(b_1 \vee b_2) H(p(b_1|b_1 \vee b_2), p(b_2|b_1 \vee b_2)) \end{aligned} \tag{2}$$

This refers to an experiment with $n+1$ outcomes labelled by $a_1, \ldots a_{n-1}, b_1, b_2$ and the grouping of two of these outcomes together, and is clearly different from Grouping Axiom 2 (see Fig. 2). Brukner and Zeilinger's Grouping Axiom 2 is the result of applying (2) to a coarse grained joint experiment in which we only distinguish the B outcomes of $A \wedge B$ for A outcome a_n, and Grouping Axiom 3 is the result of applying the simple grouping axiom to the fine grained joint experiment n times. These are not, then, expressions of the grouping axiom, but rather demonstrate its effect when actually applied to an already well defined joint probability distribution. The absence of certain joint probability distributions in quantum mechanics does not, however, affect the meaningfulness of the grouping axiom, because in its proper formu-

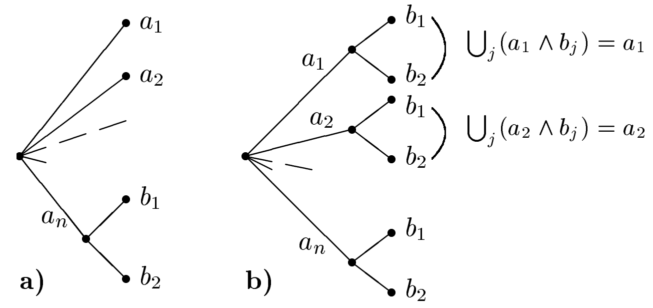

FIG. 2: **a**) in Grouping Axiom 1 and eqn. (2), the 'B' outcomes b_1 and b_2 cannot occur without $a_n \stackrel{\text{def}}{=} b_1 \cup b_2$ occurring. **b**) in the joint experiment scenario, b_1 or b_2 can occur without a_n occurring, but this is made to appear the same by coarse graining the joint experiment and only recording B values when we get A outcome a_n.

lation it does not refer to joint experiments[4] . (Note that if the outcomes of the experiment in eqn. (2) were represented by one dimensional projection operators, the event $b_1 \vee b_2$ would be represented by a sum of orthogonal projectors which commutes with the remaining projectors; a similar relation (*coexistence*) holds if the outcomes are represented by POV elements (*effects*)[22].)

Thus we see that the explicit argument fails. There may remain, however, a certain intuitive one. Brukner and Zeilinger are perhaps suggesting that we miss something importantly quantum by using the Shannon information as its derivation is restricted to the case of commuting observables. Or rather (since we have seen that the grouping axiom does not explicitly refer to joint experiments and we know that it is not in fact necessary for the functioning of the Shannon information as a measure of uncertainty anyway), because it can only tell us about the uncertainty in joint experiments for mutually commuting observables rather than the full set of observables. However, when we recall that a measure of uncertainty is a measure of the spread of a probability distribution, we see that this simply amounts to the truism that one cannot have a measure of the spread of a joint probability distribution unless one actually has a joint probability distribution, and it in no way implies that there is anything un-quantum about the Shannon information itself.

The fact that joint probability distributions cannot be defined for all possible groupings of experiments that we might consider does not tell us anything about whether a certain quantity is a good or bad measure of uncertainty for probability distributions that *can* be defined (e.g. any probability distribution derived from a quantum state by the trace rule, or conditional probabilities given by the Lüders rule). We must be careful not to confuse the question of what makes a good measure of uncertainty with the question of when a joint probability distribution can be defined.

We already know that a function of a joint probability distribution cannot be a way of telling us how much we know, or how uncertain we are, *in general* when we know the state of a system, because we know that a joint probability distribution for all possible measurements does not exist. It is for this reason in quantum mechanics that we introduce measures of mixedness such as the von Neumann entropy, which are functions of the *state* rather than of a probability distribution. It is not a failing of the Shannon information as a measure of uncertainty or expected information gain that it does not fulfill the same role. Part of Brukner and Zeilinger's worry about the Shannon information thus seems to arise because they are trying to treat it too much like a measure of mixedness, a measure of how uncertain we are in general when we know the state of a quantum system[5].

IV. BRUKNER AND ZEILINGER'S 'TOTAL INFORMATION CONTENT'

The final argument proposed against the Shannon information is that it is not appropriately related to a notion of 'total information content' for a quantum system. It is also suggested that the von Neumann entropy, which has a natural relation to the Shannon information, is not a measure of information content as it makes no explicit reference to information gain from measurements in general [5, 10].

In place of the Shannon information, Brukner and

[4] To see that the standard case and the joint experiment case are mathematically distinct, note that the joint experiment formalism cannot express the situtation in which B events only happen if the a_n event occurs. For that we would require that $p(a_n) = p(b_1) + p(b_2)$, but then the marginal distribution for B outcomes in the joint experiment does not sum to unity as is required for a well defined joint experiment, $\sum_i p(b_i) = p(a_n) \neq 1$ in general.

[5] This is illustrated for example in their reply to criticism of their grouping axiom argument by Hall [10, 23]. Hall presents an interpretation of the grouping axiom concerning the increase in randomness on mixing of non-overlapping distributions, to which Brukner and Zeilinger's worries about joint experiments would not apply. Their reply, in essence, is that the density matrix cannot be simultaneously diagonal in non-commuting bases, therefore it cannot be thought to be composed of non-overlapping classical distributions, hence Hall's grouping axiom will not apply, further supporting their original claim that the Shannon measure is tied to the notion of classical properties. What this reply in fact establishes, however, is that Hall's axiom applied to mixtures of *classical* distributions is not relevant to characterising the randomness of the *density matrix*; but this is something with which everyone would agree, and this job certainly not one for which the Shannon information is intended. (If we did wish to use the grouping axiom in characterising the randomness of the density matrix, we would apply Hall's version to mixtures of density operators with orthogonal support; this would then pick out the von Neumann entropy up to a constant factor [24].)

Zeilinger propose the quantity

$$I(\vec{p}) = \mathcal{N} \sum_{i=1}^{n} \left(p_i - \frac{1}{n} \right)^2 , \tag{3}$$

from which they derive their notion of total information content as follows:

A set of measurements is called *mutually unbiased* or *complementary* [25] if the sets of projectors $\{P\}, \{Q\}$ associated with any pair of measurement bases satisfy $\mathrm{Tr}(PQ) = 1/n$, where n is the dimensionality of the system. There can exist at most $n+1$ such bases [26], constituting a *complete* set, and as was shown by Ivanovic [27], measurement of such a complete set on an ensemble of similarly prepared systems determines their density matrix ρ completely. In analogy to acquiring the information associated with a (pointlike) classical system by learning its state (determining its position in phase space), Brukner and Zeilinger then suggest that the *total information content* of a quantum system should be given by a sum of information measures for a complete set of mutually unbiased measurements. Taking $I(\vec{p})$ as the measure of information, the result is a unitarily invariant quantity:

$$I_{tot} = \sum_{j=1}^{n+1} I(\vec{p^j}) = \sum_{ji} \left(p_i^j - \frac{1}{n} \right)^2 = \mathrm{Tr} \left(\rho - \frac{\mathbb{1}}{n} \right)^2 . \tag{4}$$

This also provides their argument against the Shannon information. It is a necessary constraint on a measure of total information content, they argue, that it be unitarily invariant, but substituting $H(\vec{p})$ for $I(\vec{p})$ in (4) does not result in a unitarily invariant quantity, that is, we do not have a sum to a 'total information content'. Let us call the requirement that a measure sum to a unitarily invariant quantity for a complete set of mutually unbiased measurements the 'total information constraint'. The suggestion is that the Shannon measure is inadequate as a measure of information gain because it does not satisfy the total information constraint and hence does not tell us how much of the total information content of a system we learn by performing measurements in a given basis. Similarly, the complaint against the von Neumann entropy is that it is merely a measure of mixedness, as unlike I_{tot}, it has no relation to the information gained in a measurement unless we happen to measure in the eigenbasis of ρ.

A few remarks are in order. First, $I(\vec{p})$ and I_{tot} are not unfamiliar expressions. The quantity $\sum_i (p_i - 1/n)^2$ is one of the class of measures of the concentration of a probability distribution given by Uffink [12], and Fano, for example [28], has remarked that $\mathrm{Tr}(\rho^2)$ can serve as a good measure of information, giving its expression in terms of orthogonal Hermitian operators (an expression very close to eqn. (4)). Note also that $I(\vec{p})$ is an *increasing* function of the concentration of a probability distribution, hence a measure of *how much we know* given a probability distribution, rather than being a measure of uncertainty like $H(\vec{p})$; similarly I_{tot} is an *increasing* function of the purity of ρ.

More importantly, however, 'information content' might mean several different things. It may not, then, be reasonable to require that every meaningful information measure sum to a unitarily invariant quantity that can be interpreted as an information content. Moreover, we may well ask why information measures for a complete set of mutually unbiased measurements should be expected to sum to *any* particularly interesting quantity. That the measure $I(\vec{p})$ happens to sum to a unitarily invariant quantity is, as we shall presently see, the consequence of a geometric property tangential to its role as a measure of information.

A. Some Different Notions of Information Content

It is useful to distinguish between the information encoded in a system, the information required to specify the state of a system (more precisely, the information required to specify a *sequence* of states drawn from a given ensemble) and states of complete and less than complete knowledge or information. Each of these can serve as a notion of information content in an appropriate context. In the classical case, their differences can be largely ignored, but in the quantum case there are important divergences. It is, for instance, necessary to introduce the concept of the accessible information to characterise the difference between information encoded and specification information [19].

If we consider encoding the outputs a_i of a classical information source A into pure states $|a_i\rangle$ of an ensemble of quantum systems, then the state of the ensemble will be given by $\rho = \sum_i p(a_i)|a_i\rangle\langle a_i|$. The von Neumann entropy, $S(\rho) = -\mathrm{Tr}\rho \log \rho$, is a measure of how mixed this state is, giving us one sense of information content - the more mixed a state, the less information we have about what the outcome of measurements on systems described by the state will be[6].

If we are presented with a sequence of systems drawn from an ensemble prepared in this manner, each will be in one of the pure states, and the number of bits per system required to specify this sequence will be $H(A)$, the information of the classical source (which will be greater than $S(\rho)$ unless we have coded in orthogonal states). This is the *specification information*, also the amount of information required to prepare the sequence. For the amount of information that has actually been encoded

[6] Mixed states are also sometimes said to be states of less than complete information due to a lack of information about the way a system was prepared, represented by a probability distibution over possible pure states. Our reading is to be preferred given the many-one relation of preparation procedures to density operators and the fact that density operators can also result from tracing out unwanted degrees of freedom.

into the systems, however, we need to consider measurements on the ensemble and the Shannon mutual information $H(A:B)$.

As already remarked (Section III A), if we encode using a certain basis (our $|a_i\rangle$ form an orthonormal set) and we measure in a different basis, then $H(A:B) < H(A)$; quantum 'noise' reduces the amount we can decode. More significantly, if we have coded in non-orthogonal states (if, for example the number of outputs of our classical source is greater than the dimensionality of our quantum systems), then no measurement can distinguish these states perfectly and we cannot recover the complete classical information. Taking 'encoded' to be a 'success' word (something cannot be said to have been encoded if it cannot in principle be decoded), then the maximum amount of information encoded in a system is given by the *accessible information*, the maximum over all decoding observables of the mutual information. Using non-orthogonal coding states, the amount we can encode is less than the specification information. A well known result due to Holevo [29] provides an upper bound on the mutual information resulting from measurement of any observable (including POV measurements). For the general case of encoding states ρ_{a_i}, this is given by

$$H(A:B) \leq S(\rho) - \sum_i p(a_i)S(\rho_{a_i}),$$

which in the case we are considering of pure encoding states, reduces to $H(A:B) \leq S(\rho)$. This provides a very strong sense in which the von Neumann entropy does give us a notion of the total information content of a quantum system - it is the maximum amount that can actually be encoded in the system.

Brukner and Zeilinger do not consider a quantum communication channel but are concerned rather with the information content of a single system considered in isolation. This information content is supposed to relate to how much we learn from learning the state, but if the system is being treated in isolation then by learning its state we are not acquiring a certain amount of information in virtue of the state being drawn from a given ensemble, as in the standard notion of information. (Hence their analogy with gaining the information content of a classical system fails to hold.) In fact, their 'total information content' seems best interpreted as a measure of mixedness analogous to the von Neumann entropy.

When introduced [8], the information measure $I(\vec{p})$ is presented as a measure of how much we know about what the outcome of a particular experiment will be, given the state. The *total* information of the state, then, would seem to be a measure of how much we know *in general* about what the outcomes of experiments will be given the state; and this is precisely a question of the degree of mixedness of the state[7].

1. Measures of mixedness

The functioning of measures of mixedness can usefully be approached via the notions of majorization and Schur convexity (concavity). The majorization relation $\prec$ imposes a pre-order on probability distributions [12, 14]. A probability distribution $\vec{q}$ is majorized by $\vec{p}$, $\vec{q} \prec \vec{p}$, iff $q_i = \sum_j S_{ij}p_j$, where S_{ij} is a doubly stochastic matrix. That is (via Birkhoff's theorem), if $\vec{q}$ is a mixture of permutations of $\vec{p}$. Thus if $\vec{q} \prec \vec{p}$, then $\vec{q}$ is a more mixed or disordered distribution than $\vec{p}$.

Schur convex (concave) functions respect the ordering of the majorization relation: a function f is Schur convex if, if $\vec{q} \prec \vec{p}$ then $f(\vec{q}) \leq f(\vec{p})$, and Schur concave if, if $\vec{q} \prec \vec{p}$ then $f(\vec{q}) \geq f(\vec{p})$ (for strictly Schur convex(cave) functions, equality holds only if $\vec{q}$ and $\vec{p}$ are permutations of one another). This explains the utility of such functions as measures of the concentration and uncertainty of probability distributions, respectively.

The majorization relation will apply equally to the vectors of eigenvalues of density matrices. It can be shown that the vector of eigenvalues $\vec{\lambda}'$ of the density matrix ρ' of the post measurement ensemble for a (non-selective) projective measurement is majorized by the vector of eigenvalues $\vec{\lambda}$ of the pre-measurement state ρ [14]. (If we measure in the eigenbasis of ρ, then there is, of course, no change in the eigenvalues). The λ'_i are just the probabilities of the different outcomes of the measurement in question, thus the probability distribution for the outcomes of any given measurement will be more disordered or spread than the eigenvalues of ρ.

If we take any Schur concave function we know to be a measure of uncertainty, for instance the Shannon information $H(\vec{p})$, and $\vec{p}$ is the probability distribution for measurement outcomes, we then know that $H(\vec{p}) \geq H(\vec{\lambda})$, for any projective measurement we might perform. This explains why $H(\vec{\lambda}) = S(\rho)$, a measure of mixedness, is a measure of how much we know given the state: the more mixed a state, the more uncertain we must be about the outcome of any given measurement. Similarly, if we take a measure of the concentration of a probability distribution, a Schur convex function such as $I(\vec{p})$, then we know that for any measurement with outcome probability distribution $\vec{p}$, $I(\vec{p}) \leq I(\vec{\lambda}) = I_{tot}$; and this explains why I_{tot} is a measure of how much we know given ρ: the less the value of I_{tot}, the less able we are to predict the outcome of any given experiment.

[7] Recently it has been noted that I_{tot} is also related to the average distance of our estimate of the unknown state from the true state (measured in the Hilbert-Schmidt norm), given only a finite number N of experiments in each mutually unbiased basis [30]. This seems best understood as indicating that the mean error in our state estimation is inversely related to N, with a constant of proportionality that depends on the dimension of the system and the mixedness of the state. In any case, Brukner and Zeilinger are primarily interested in how much we know when the state has been determined to arbitrary accuracy.

Brukner and Zeilinger would of course deny that their total information content is merely a measure of mixedness. The argument that it is more than this rests on the satisfaction of the total information constraint, the relation between the measure of information $I(\vec{p})$ and I_{tot} for a complete set of mutually unbiased measurements as expressed in eqn. (4). We shall now see that this relation can be given a simple geometric explanation using the Hilbert-Schmidt representation of density operators.

B. The Relation between Total Information Content and $I(\vec{p})$

The set of complex $n \times n$ Hermitean matrices forms an n^2-dimensional real Hilbert space $V_h(C^n)$ on which we have defined an inner product $(A, B) = \mathrm{Tr}(AB); A, B \in V_h(C^n)$ and a norm $\|A\| = \sqrt{\mathrm{Tr}(A^2)}$ [28, 31]. The density matrix ρ of an n dimensional quantum system can be represented as a vector in this space. The requirements on ρ of unit trace and positivity imply that the tip of any such vector must lie in the $n^2 - 1$ dimensional hyperplane $\boldsymbol{T}$ a distance $1/\sqrt{n}$ from the origin and perpendicular to the unit operator $\mathbb{1}$, and on or within a hypersphere of radius one centred on the origin.

It is useful to introduce a set of basis operators on our space; we require n^2 linearly independent operators $U_i \in V_h(C^n)$ and it may be useful to require orthogonality: $\mathrm{Tr}(U_i U_j) = \text{const.} \times \delta_{ij}$. Any operator on the system can then be expanded in terms of this basis and in particular, ρ can be written as

$$\rho = \frac{\mathbb{1}}{n} + \sum_{i=1}^{n^2-1} \mathrm{Tr}(\rho U_i) U_i,$$

where we have chosen $U_0 = \mathbb{1}$ to take care of the trace condition.

Evidently, ρ may be determined experimentally by finding the expectation values of the $n^2 - 1$ operators U_i in the state ρ. If we include the operator $\mathbb{1}$ in our basis set, then the idempotent projectors associated with measurement of any maximal (non-degenerate) observable will provide a maximum of a further $n - 1$ linearly independent operators. Obtaining the probability distribution for a given maximal observable will thus provide $n - 1$ of the parameters required to determine the state, and the minimum number of measurements of maximal observables that will be needed in total is $n + 1$, if each observable provides a full complement of linearly independent projectors.

Each such set of projectors spans an $n - 1$ dimensional hyperplane in $V_h(C^n)$ and their expectation values specify the projection of the state ρ into this hyperplane. Ivanovic [27] noted that projectors P, Q belonging to any two different mutually unbiased bases will be orthogonal in $\boldsymbol{T}$, hence the hyperplanes associated with measurement of mutually unbiased observables are orthogonal in the space in which density operators are constrained to lie in virtue of the trace condition. If $n + 1$ mutually unbiased observables can be found, then, $V_h(C^n)$ can be decomposed into orthogonal subspaces given by the one dimensional subspace spanned by $\mathbb{1}$ and the $n + 1$ subspaces associated with the mutually unbiased observables. The state ρ can then be expressed as:

$$\rho = \frac{\mathbb{1}}{n} + \sum_{j=1}^{n+1} \sum_{i=1}^{n} q_i^j \bar{P}_i^j, \tag{5}$$

where $\bar{P}_i^j = P_i^j - \mathbb{1}/n$ is the projection onto $\boldsymbol{T}$ of the ith idempotent projector in the jth mutually unbiased basis set, and $q_i^j = (p_i^j - 1/n)$ is the expectation value of this operator in the state ρ. For a given value of j, the vectors $\bar{P}_i$ span an $(n-1)$ dimensional orthogonal subspace and the square of the length of a vector expressed in the form (5) lying in subspace j will be given by $\sum_{i=1}^{n} (q_i^j)^2 = I(\vec{p^j})$.

The geometrical explanation of I_{tot} is then simply as follows. $\mathrm{Tr}(\rho^2)$ is the square of the length of ρ in $V_h(C^n)$ and $\mathrm{Tr}(\rho - \mathbb{1}/n)^2$ is the square of the distance of ρ from the maximally mixed state (the length squared of ρ in $\boldsymbol{T}$). This squared length will just be the sum of the squares of the lengths of the components of the vector $\rho - \mathbb{1}/n$ in the orthogonal subspaces into which we have decomposed $V_h(C^n)$, i.e. it will be given by $\sum_{ji} (q_i^j)^2$. This is what eqn. (4) reports and it explains how I_{tot} and $I(\vec{p})$ satisfy the total information constraint.

Thus we see that if I_{tot} differs from being a simple measure of mixedness, then that is because it is a measure of length also; and this explains why it can be given by a sum of quantities $I(\vec{p})$ for a complete set of mutually unbiased measurements. As measures of *how much we know given the state*, however, I_{tot} and $S(\rho)$ bear the same relation to their appropriate measures of information, as we saw in the previous section. Equally, as measures of information, $H(\vec{p})$ stands to $S(\rho)$ in the same relation as $I(\vec{p})$ to I_{tot}. I_{tot} is the *upper* bound on the amount we can *know* about the outcome of a measurement as measured by $I(\vec{p})$; $S(\rho)$ is the *lower* bound on our *uncertainty* about what the outcome of a measurement will be, as measured by $H(\vec{p})$.

The complaint against the Shannon information was supposed to be that as $H(\vec{p})$ fails to satisfy the total information constraint, it does not tell us the information gained from a particular measurement; the complaint against the von Neumann entropy that as $S(\rho)$ is not given by a sum of measures for a complete set of mutually unbiased measurements, it is not suitably related to the information gained from a measurement. However, we can now see that insisting on the total information constraint in this way is tantamount to insisting that only a function which measures the length of the component of ρ lying in a given hyperplane can be a measure of information, and correlatively, that the only viable notion of total information content is a measure of the length of ρ in $V_h(C^n)$. But $H(\vec{p})$ can be a perfectly good measure of information without having to be a measure of the length of the projection of ρ into the subspace associated with an observable; and as we have just seen, $S(\rho)$ does have an explicit relation to the information gain from measurement that justifies its interpretation as a total information content. A relation, moreover, that I_{tot} also possesses and which serves to justify *its* interpretation as a measure of how much we know given the state.

Hence our conclusion must be that the total information constraint is not a resonable requirement on measures of information.

Of course, $H(\vec{p})$ does not tell us the information gain on measurement it we take, as Brukner and Zeilinger seem to, 'the information encoded in a basis' simply to *mean* the length squared of the component of the state lying in the measurement hyperplane; but this is a non-standard usage. $H(\vec{p})$ certainly remains a measure of our expected information gain from performing a particular measurement (how much the outcome will surprise us, on average, given that we have the probability distribution); and if we are interested in the amount of information encoded, in the usual sense of the word, that can be decoded using a particular measurement, i.e. if we have a string of systems into which information has actually been encoded, then we may always just consider the Shannon mutual information associated with that measurement. (The 'total information' associated with this quantity will then be given, via the Holevo bound, as the von Neumann entropy.)

V. CONCLUSION

Of the three arguments that Brukner and Zeilinger have presented against the Shannon information, we have seen that the first two fail outright. These arguments sought to establish that the notion of the Shannon information is undermined in the quantum context due to a reliance on classical concepts. With regard to the first we saw that, contrary to Brukner and Zeilinger, the existence of a pre-determined string is neither necessary nor sufficient for the interpretation of $H(\vec{p})$ as a measure of information, hence the absence of such a string would not cause any problems for the Shannon information in quantum mechanics.

The objective of their second argument was to highlight classical assumptions in the grouping axiom that would prevent the axiomatic derivation of the Shannon information going through in the quantum case. This argument turned out to be based on an erroneous reading of the grouping axiom that appeals to joint experiments. The grouping axiom is in fact perfectly well defined in the quantum case and the standard axiomatic derivation of the form of $H(\vec{p})$ can indeed go through. The grouping axiom does not reveal any problematic classical assumptions implicit in the Shannon information.

In their final argument, Brukner and Zeilinger suggest that defining the notion of the total information content of a quantum system in terms of the Shannon information would lead to a quantity with the unnatural property of unitary non-invariance. But this is not a compelling argument against the Shannon quantity as a measure of information, however. We have seen that it is not a necessary requirement on every meaningful measure of information that it sum to a unitarily invariant quantity for a complete set of mutually unbiased measurements; nor, conversely, is it necessary that every viable notion of total information content be given by such a sum of individual measures of information.

Brukner and Zeilinger's arguments thus fail to establish that the Shannon information involves any particularly classical assumptions or that there is any difficulty in the application of the Shannon measure to measurements on quantum systems. The Shannon information is perfectly well defined and appropriate as a measure of information in the quantum context as well as in the classical.

Acknowledgments

Thanks are due to Harvey Brown for useful discussion and in particular, for emphasis of the point raised in footnote (1); thanks also to Jos Uffink and Michael Hall for comments. This work was supported by a studentship from the UK Arts and Humanities Research Board.

[1] C. Fuchs (2001), arXiv:quant-ph/0106166.
[2] N. Mermin, in *Quantum (Un)speakables: Essays in Commemoration of John S. Bell*, edited by R. Bertlmann and A. Zeilinger (Springer-Verlag, Berlin, Heidleberg, 2001), forthcoming, arXiv:quant-ph/0107151.
[3] J. Wheeler, in *Complexity, Entropy and the Physics of Information*, edited by W. Zurek (Addison-Wesley, Redwood City, CA, 1990), pp. 3–28.
[4] A. Zeilinger, Found. Phys. **29**, 631 (1999).
[5] C. Brukner and A. Zeilinger, Phys. Rev. A **63**, 022113 (2001).
[6] C. Shannon, Bell Syst. Tech. J. **27** (1948), repr. in [7] pp.30-125.
[7] C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana and Chicago, 1963), Illini Press ed.
[8] C. Brukner and A. Zeilinger, Phys. Rev. Lett. **83**, 3354 (1999).
[9] C. Brukner and A. Zeilinger, J. Mod. Opt. **47**, 2233 (2000).
[10] C. Brukner and A. Zeilinger (2000), arXiv:quant-ph/0008091.
[11] C. Timpson, B.Phil. Thesis, University of Oxford (2000), URL `http://users.ox.ac.uk/~quee0776/thesis.html`.

[12] J. Uffink, Ph.D. thesis, University of Utrecht (1990).
[13] H. Maassen and J. Uffink, Phys. Rev. Lett. **60**, 1103 (1988).
[14] M. Nielsen, Phys. Rev. A **63**, 022114 (2001).
[15] R. Ash, *Information Theory* (Interscience Publishers, 1965).
[16] J. Bub, *Interpreting the Quantum World* (Cambridge University Press, 1997).
[17] J. Bell, Found. Phys. **12**, 989 (1982).
[18] A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer Academic Publishers, Dordrecht, 1995).
[19] B. Schumacher, Phys. Rev. A **51**, 2378 (1995).
[20] E. Jaynes, Phys. Rev. **106**, 620 (1957).
[21] D. Faddeev, in *Arbeiten zum Informationstheorie I*, edited by H. Grell (Deutscher Verlag der Wissenschaften, Berlin, 1957), pp. 88–91.
[22] P. Busch, P. J. Lahti, and P. Mittelstaedt, *The Quantum Theory of Measurement* (Springer-Verlag, Berlin, Heidelberg, 1996), 2nd ed.
[23] M. J. Hall (2000), arXiv:quant-ph/0007116.
[24] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).
[25] J. Schwinger, Proc. Nat. Acad. Sci. USA **46**, 570 (1960).
[26] W. K. Wootters and B. D. Fields, Ann. Phys. (NY) **191**, 363 (1989).
[27] I. Ivanovic, J. Phys. A **14**, 3241 (1981).
[28] U. Fano, Rev. Mod. Phys. **29**, 74 (1957).
[29] A. Holevo, Probl. Inf. Transm. (USSR) **9**, 177 (1973).
[30] J. Rehacek and Z. Hradil (2001), arXiv:quant-ph/0111115.
[31] E. Wichmann, J. Math. Phys. **4**, 884 (1963).