The role of relative entropy in quantum information theory

V. Vedral*

Centre for Quantum Computation, Clarendon Laboratory, University of Oxford, OX1 3PU, United Kingdom

(Published 8 March 2002)

Quantum mechanics and information theory are among the most important scientific discoveries of the last century. Although these two areas initially developed separately, it has emerged that they are in fact intimately related. In this review the author shows how quantum information theory extends traditional information theory by exploring the limits imposed by quantum, rather than classical, mechanics on information storage and transmission. The derivation of many key results differentiates this review from the usual presentation in that they are shown to follow logically from one crucial property of relative entropy. Within the review, optimal bounds on the enhanced speed that quantum computers can achieve over their classical counterparts are outlined using information-theoretic arguments. In addition, important implications of quantum information theory for thermodynamics and quantum measurement are intermittently discussed. A number of simple examples and derivations, including quantum superdense coding, quantum teleportation, and Deutsch's and Grover's algorithms, are also included.

CONTENTS

I. INTRODUCTION

Quantum physics not only provides the most complete description of physical phenomena known to man, it also provides a new philosophical framework for our understanding of nature. It enables us to accurately model systems ranging in size from quarks and atoms to large cosmic objects such as black holes. Information theory, on the other hand, teaches us about our physical ability to store and process information. Without a formalized information theory, many of the recent developments in telecommunications, computer science, and engineering would simply not have been possible. Although quantum physics and information theory initially developed separately, their recent integration is seen as yet another important step towards understanding the fundamental properties and limitations of Nature.

One of the central information-theoretic concepts in science is that of distinguishability. Inevitably an animal's survival depends on its ability to distinguish a mate from a predator or prey. In the same way, physical experiments aim to be sensitive enough to be able to distinguish one hypothesis from another. It is, however, no surprise that the influence of the concept of distinguishability is felt far beyond science. Life consists of a series of decisions that have to be made. This we do, consciously or unconsciously, by evaluating all the alternatives and distinguishing the consequences of various alternative actions.

The purpose of this review is to show that the apparently simple concept of distinguishability is at the root of information processing. Ultimately how well we can distinguish different physical states determines how much information we can encode into a certain system and how quickly we can manipulate it. Distinguishability in turn is completely dependent upon the laws of physics, and quantum physics naturally allows for more versatile information processing than does classical physics. The reasoning behind this is that unlike classical states, two

*Permanent address; Blackett Laboratory, Imperial College, London SW7 2BZ.

different quantum states are not necessarily fully distinguishable. It is interesting to note that although this at first seems like a limitation, it in fact presents us with significantly more possibilities for information encoding and transmission.

In this review I first plan to argue that relative entropy is the most appropriate quantity for measuring distinguishability between different states. The proper framework in which to talk about states is, of course, quantum mechanics, so it is necessary to define quantum relative entropy. I prove that relative entropy, both classical and quantum, does not increase with time. Thus two states can only become less distinguishable as they undergo any kind of evolution. This result will be central to my review, as subsequent results will follow from this simple fact.

I then go on to show that the "no increase of relative entropy" principle tells us about the ability of quantum states to store and process information. Information has to be encoded and manipulated in physical systems. Therefore distinguishability of different states within a physical system is a prerequisite. Looking at this from the point of view of communication, what does it mean to send and receive a message? Sending a message successfully means encoding the information we wish to send into a structured format that the receiver must be able to distinguish unambiguously. Communication capacity can then be thought of as the rate at which we can send and receive messages. The rate of successful transmission is determined by the relative entropy between various encoding states.

What is less obvious, but nonetheless equally true, is that computation can also be viewed as a special kind of communication. This will allow the use of relative entropy to quantify the efficiency (i.e., speed) of quantum computation in general.

The role of measurement within quantum mechanics and therefore information theory is paramount. Classically the measurement process is implicit because physical quantities have well-defined preexisting properties. For example, a classical bit is either in the state 0 or 1, whereas a quantum bit can exist in a combination of the two states. A measurement is necessary to "collapse" this combination to a classical result that we can then read. The very concept of measurement efficiency can also be quantified using relative entropy. A measurement, like a communication process, creates correlations between a system and an apparatus whose purpose is to receive an amount of information from the system. The opposite of this process, namely, the deletion of information, can be seen to be at the root of irreversibility, and this invariably contributes to an increase in the entropy of the environment. This amount is exactly quantified using the relative entropy between the environmental state and the apparatus state, and it provides an exciting link between information theory, computation, thermodynamics, and quantum mechanics. However, before we reach this exciting stage, our long journey has to begin with a much simpler question: how do we quantify uncertainty in a physical state?

II. RELATIVE ENTROPY

Fundamental to our understanding of distinguishability is the measure of *uncertainty* in a given probability distribution. This uncertainty can be quantified by introducing the idea of "surprise." Suppose that a certain event happens with a probability p. We would then like to quantify how surprised we are when that event does happen. The first guess would be $1/p$: the smaller the probability of an event, the more surprised we are when the event happens, and vice versa. However, an event might be composed of two independent events that happen with probabilities q and r, respectively, so that the probability of both events occurring is $p=q\times r$. We would now intuitively expect that the surprise of p is the same as the surprise of q plus the surprise of r. But $1/p\neq 1/q+1/r$, so that $1/p$ is not really a satisfactory definition from this perspective. Instead, if we define surprise as $\ln(1/p)$, then the above property called *additivity* is satisfied, since $-\ln p_1p_2=-\ln p_1-\ln p_2$. With a probability distribution $\Sigma_n p_n=1$, the total uncertainty is just the average of all the surprises. Additivity of uncertainties of statistically independent events is such a stringent condition that it basically leads to a unique measure (Shannon and Weaver, 1949) up to a constant and logarithmic base.

Definition. The uncertainty in a collection of possible states a_i with corresponding probability distribution $p(a_i)$ is given by its *entropy*,

$$S(p):=-\sum_i p(a_i)\ln p(a_i), \tag{1}$$

called the *Shannon entropy*. We note that there is no Boltzmann constant term in this expression as there is for the physical entropy, since it is by convention set to unity. This measure is suitable for the states of systems described by the laws of classical physics, but it will have to be changed, along with other classical measures, when we present the quantum information theory.

We ultimately wish to be able to talk about storing and processing information. For this we require a means of comparing two different probability distributions, which is why I introduce the notion of *relative entropy* (first introduced by Kullback and Leibler, 1951). Suppose that a collection of events has the probability distribution $\{p_i\}$, but we mistakenly think that this probability distribution is $\{q_i\}$. For example, we have a coin that we think is fair, i.e., the probability of getting a head or a tail when the coin is tossed is equal. If we toss this coin n times, on average we expect heads half of the time and tails the other half. In reality, the coin, by virtue of its uneven weight distribution, will not be completely fair, so our expectation will turn out to be wrong. There will consequently be a discrepancy between our expected and real probability distribution. This discrepancy is very frequently the case in real life, and it is, in fact, very rare that we have complete information about any event. Therefore we can formalize that when a par-

ticular outcome j happens, we associate surprise $-\ln q_j$ with it. The average surprise, or information, according to this erroneous belief, is

$$-\sum_i p_i \ln q_i .$$

Since events happen with probabilities $\{p_i\}$ (in spite of our belief), these are the correct ones to feature in the averaging process. However, the real amount of information we are obtaining is, as defined earlier, given by the Shannon entropy $S(p)=-\Sigma_i p_i \ln p_i$. It is not so difficult to show that $S(p)\leqslant -\Sigma_i p_i \ln q_i$ (equality holds if and only if $p_i=q_i$ for all i) so that there is an "uncertainty deficit," as it were, stemming from our wrong assumption and equal to the difference between the two averages. This deficit quantity is called the relative entropy.

Definition. Suppose that we have two sets of discrete events a_i and b_j with the corresponding probability distributions, $p(a_i)$ and $p(b_j)$. The *relative entropy* between these two distributions is defined as

$$S[p(a)\|p(b)] := \sum_i p(a_i) \ln \frac{p(a_i)}{p(b_i)}. \tag{2}$$

This function is a measure of the "distance" between $p(a_i)$ and $p(b_j)$, even though, strictly speaking, it is not a mathematical metric since it fails to be symmetric: $S[p(a)||p(b)]\neq S[p(b)||p(a)]$. This is interesting, since at first it looks as if there should be no difference between mistaking the probability distribution p_i for q_i, or vice versa. Intuitively this can be explained using our coin example. Suppose that someone gives us a coin that is either fair or completely unfair, e.g., it always gives heads. Now we have to toss this coin a number of times and infer which of the two coins we have. If we toss the fair coin and obtain tails, then our inference will immediately be that we have the fair coin. If, however, we obtain heads, then it could be either coin. If we tossed it more times, the fair coin would eventually give us tails. If, however, we were holding the completely unfair coin from the beginning, then even after 100 heads we could never really eliminate the possibility that it is the fair coin, since this outcome is statistically possible (although highly unlikely). Therefore how certain we are about which coin we hold is clearly dependent on whichever coin we hold and how different it is from the other one. As we shall see shortly, our uncertainty is quantified by the relative entropy, and it is thus to be expected that it is asymmetric. I now describe this statistical approach in more detail.

A. Statistical significance

A more operational interpretation of both the Shannon entropy and the relative entropy comes from the statistical point of view. The generalization of this formalism to the quantum domain will be presented in the next section and I shall offer an operational interpretation of the measures of quantum correlations to be introduced therein. I now follow the approaches of Cover and Thomas (1991), and Csiszár and Körner (1981); the reader interested in more detail should consult these two books.

Let $X_1,X_2,\ldots,X_n$ be a sequence of n symbols from an alphabet $A=\{a_1,a_2,\ldots,a_{|A|}\}$, where $|A|$ is the size of the alphabet. We denote a sequence $x_1,x_2,\ldots,x_n$ by x^n or, equivalently, by $\mathbf{x}$. The type $P_{\mathbf{x}}$ of a sequence $x_1,x_2,\ldots,x_n$ will be called the relative proportion of occurrences of each symbol of A, i.e., $P_{\mathbf{x}}(a)=N(a|\mathbf{x})/n$ for all $a\in A$, where $N(a|\mathbf{x})$ is the number of times the symbol a occurs in the sequence $\mathbf{x}\in A^n$. Thus, according to this definition, the sequences 011010 and 100110 are of the same type. $\mathcal{P}_n$ will denote the set of types with denominator n. If $P\in\mathcal{P}_n$, then the set of sequences of length n and type P is called the *type class* of P, denoted by $T(P)$, i.e., mathematically

$$T(P)=\{\mathbf{x}\in A^n : P_{\mathbf{x}}=P\}.$$

We now approach the first theorem about types, which is at the heart of the success of this theory and states that the number of types increases only polynomially with n.

Theorem 1.

$$|\mathcal{P}_n|\leqslant (n+1)^{|A|}.$$

Proof of this is left for the reader, but the rationale is simple. Suppose that we generate an n string of 0's and 1's. The number of different types is then $n+1$, i.e., polynomial in n: the zeroth type has only one string—all zeros; the first type has n strings—all strings containing exactly one 1; the second type has $n(n-1)/2$ strings—all those containing exactly two 1's, and so on; the nth type has only one sequence—all ones. The most important point is that the number of sequences is exponential in n, so that at least one type has exponentially many sequences in its type class, since there are only polynomially many different types. A simple example is a coin tossed n times. If it is a fair coin, then we expect heads half of the time and tails the other half of the time. The number of all possible sequences for this coin is 2^n (i.e., exponential in n) where each sequence is equally likely (with probability 2^{-n}). However, the size of the type class in which there is an equal number of heads and tails is $C^n_{n/2}$ (the number of possible ways of choosing $n/2$ element out of n elements), the log of which tends to n for large n. Hence this type class is in some sense asymptotically as large as all the type classes together.

We now arrive at a very important theorem that, in fact, presents the basis of the statistical interpretation of the Shannon entropy and relative entropy.

Theorem 2. If $X_1,X_2,\ldots,X_n$ are drawn according to $Q(x)$, then the probability of $\mathbf{x}$ depends only on its type and is given by

$$Q^n(\mathbf{x})=e^{-n[S(P_{\mathbf{x}})+S(P_{\mathbf{x}}\|Q)]}.$$

Proof.

$$Q^n(\mathbf{x})=\prod_{i=1}^{n} Q(x_i)=\prod_{a\in A} Q(a)^{N(a|\mathbf{x})}$$

$$=\prod_{a\in A} Q(a)^{nP_{\mathbf{x}}(a)}$$

$$=\prod_{a\in A} e^{nP_{\mathbf{x}}(a)\ln Q(a)}$$

$$=\exp\left\{n\sum_{a\in A}-P_{\mathbf{x}}(a)\ln\frac{P_{\mathbf{x}}(a)}{Q(a)}+P_{\mathbf{x}}(a)\ln P_{\mathbf{x}}(a)\right\}$$

$$=e^{-n[S(P_{\mathbf{x}})+S(P_{\mathbf{x}}||Q)]}. \quad \blacksquare$$

Therefore the probability of a sequence becomes exponentially small as n increases. Indeed, our coin-tossing example shows this: a probability for any particular sequence (such as, for example, 0000011111) is 2^{-n}. (Note: the reason that we are using e in our theorems instead of 2 is that we are also using ln instead of log.) This is explicitly stated in the following corollary.

Corollary. If $\mathbf{x}$ is the type class of Q, then

$$Q^n(\mathbf{x})=e^{-nS(Q)}.$$

The proof is left to the reader.

As n gets large, most of the sequences become typical and are all equally likely. Therefore the probability of every typical sequence times the number of typical sequences has to be equal to unity in order to conserve total probability ($e^{-nS(Q)}N=1$). From this we can see that the number of typical sequences is $N=e^{nS(Q)}$ (we consider this point more formally below). Hence the above theorem has very important implications in the theory of statistical inference and the distinguishability of probability distributions. To see how this comes about we state two theorems that give bounds on the size and probability of a particular type class. The proofs follow directly from the above two theorems and the corollary (Csiszar and Korner, 1981; Cover and Thomas, 1991).

Theorem 3. For any type $P\in\mathcal{P}_n$,

$$\frac{1}{(n+1)^{|A|}}e^{nS(P)}\leqslant|T(P)|\leqslant e^{nS(P)}.$$

This theorem provides the exact bounds on the number of typical sequences. Suppose that we have a probability distribution p_1 and p_2 for heads and tails, respectively, and we toss the coin n times. The typical (most likely) sequence will be the one in which we have p_1n heads and p_2n tails. The number of such sequences is

$$C^n_{p_1n}=\frac{n!}{(p_1n)!(p_2n)!}\sim e^{n(-p_1\ln p_1-p_2\ln p_2)},$$

i.e., an exponential in n (more tosses, more possibilities) and entropy (higher uncertainty, more possibilities). The next theorem offers a statistical interpretation of the relative entropy.

Theorem 4. For any type $P\in\mathcal{P}_n$, and any distribution Q, the probability of the type class $T(P)$ under Q^n is $e^{-nS(P||Q)}$ to first order in the exponent. More precisely,

$$\frac{1}{(n+1)^{|A|}}e^{-nS(P||Q)}\leqslant Q^n[T(P)]\leqslant e^{-nS(P||Q)}.$$

The meaning of this theorem is that if we draw results according to Q, the probability that it will look as if it was drawn from P is exponentially decreasing with n and relative entropy between P and Q. The closer Q is to P, the higher the probability that their statistics will look the same. Alternatively, the higher the number of draws n, the smaller the probability that we will confuse the two. (We can see an explicit example below.) The above two results can be succinctly written in an exponential fashion that will be useful to us as

$$|T(P)|\to e^{-nS(P)}, \tag{3}$$

$$Q^n[T(P)]\to e^{-nS(P||Q)}. \tag{4}$$

The first statement also leads to the idea of *data compression*, in which a string of length n generated by a source with entropy S can be encoded into a string of length nS. The second statement says that if we are performing n experiments according to distribution Q, the probability that we will get something that looks as if it was generated by distribution P decreases exponentially with n, depending on the relative entropy between P and Q. This idea immediately leads to Sanov's theorem, whose quantum analog will provide a statistical interpretation of one measure of entanglement presented in Sec. IV. We now consider data compression and Sanov's theorem.

1. Classical data compression

Suppose that we have a binary source generating 0's with twice as great a probability as that of 1's, so that the Shannon entropy is $S=\ln 3-2/3\ln 2=0.64$. Imagine that we have a string of 15 digits coming out of this source. Then, according to the above considerations [Eq. (3)], the most likely type will be the one with ten 0's and five 1's. But the size of this class is only $0.64\times 15\approx 10$, so we can use only 10 digits to encode all the above sequences of 15 numbers just by assigning the following conventional mapping: The first sequence of 15 numbers is to be encoded in 0000000000, the second sequence is to be encoded in 0000000001, …, the e^{10}th sequence is to be encoded in 1111111111. This encoding is for obvious reasons called data compression. This, in fact, offers a statistical reason for employing the Shannon entropy as a measure of uncertainty. This result is known as Shannon's lower bound (or Shannon's First Theorem) on data compression, i.e., a message cannot be compressed per bit to less than its Shannon entropy (Shannon and Weaver, 1949). There are a number of different methods used for compression, each with varying degrees of success dependent on the statistical distribution of the message; see, for example, Cover and Thomas (1991).

Now we look at the distinguishability of two probability distributions. Suppose we would like to check whether a given coin is fair, i.e., whether it generates a head-tail distribution of $f=(1/2,1/2)$. When the coin is biased it will produce some other distribution, say uf

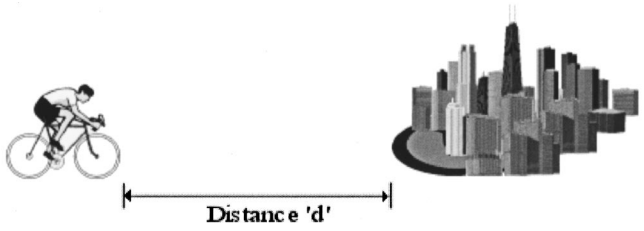

FIG. 1. The concept of distinguishability. What do we mean by the distance from the cyclist to the city in the figure? It is defined as the distance from the cyclist to the closest house in the city. Also, which distance measure is the most appropriate for measuring this? In the text I argue that when it comes to distinguishing between two or more probability distributions, the most appropriate measure is the relative entropy.

$=(1/3,2/3)$. So the question of the coin's fairness boils down to how well we can differentiate between two given probability distributions given a finite number n of experiments to perform on one of the two distributions. In the case of a coin we would toss it n times and record the number of 0's and 1's. From simple statistics (Cover and Thomas, 1991) we know that if the coin is fair then the number of 0's, $N(0)$, will be roughly $n/2-\sqrt{n} \leqslant N(0) \leqslant n/2+\sqrt{n}$, for large n, and the same for the number of 1's. Therefore if our experimentally determined values do not fall within the above limits the coin is not fair. We can look at this from another point of view, which is in the spirit of the method of types; namely, what is the probability that a fair coin will be mistaken for an unfair one with the distribution of (1/3,2/3) given n trials of the fair coin? For large n the answer is given in the previous subsection,

$$p(\text{fair} \rightarrow \text{unfair}) = e^{-nS(uf||f)},$$

where $S_{cl}(uf||f) = 1/3 \ln 1/3 + 2/3 \ln 2/3 - 1/3 \ln 1/2 - 2/3 \ln 1/2$ is the Shannon relative entropy for the two distributions. So

$$p(\text{fair} \rightarrow \text{unfair}) = 3^n 2^{-(5/3)n},$$

which tends exponentially to zero with $n \rightarrow \infty$. In fact we see that after ~ 20 trials the probability of mistaking the two distributions is vanishingly small, $<10^{-10}$. This leads to the following important result (Sanov, 1957).

2. Sanov's theorem

If we have a probability distribution Q and a set of distributions $E \subset \mathcal{P}$, then

$$Q^n(E) \rightarrow e^{-nS(P^*||Q)}, \tag{5}$$

where P^* is the distribution in E that is closest to Q using the Shannon relative entropy (see Fig. 1).

This can also be rephrased in the language of distinguishability: when we are distinguishing a given distribution from a set of distributions, then what matters is how well we can distinguish that distribution from the closest one in the set (see Fig. 1). When we turn to the quantum case later, the probability distributions will become quantum densities representing various states of a quantum system, and the question will be how well we can distinguish between these states. Note that we could also talk about Q coming from a set of states, in which case we would have $S(P||Q^*)$, where Q^* is the state that minimizes the relative entropy (i.e., the closest state).

B. Other information measures from relative entropy

Another important concept derived from relative entropy concerns the gathering of information. When one system learns something about another, their states become correlated. How correlated they are, or how much information they have about each other, can be quantified by the mutual information.

Definition. The *Shannon mutual information* between two random variables A and B, having a joint probability distribution $p(a_i,b_j)$ and therefore marginal probability distributions $p(a_i)=\Sigma_j p(a_i,b_j)$ and $p(b_j)=\Sigma_i p(a_i,b_j)$, is defined as

$$I_S(A:B) := S[p(a)] + S[p(b)] - S[p(a,b)]. \tag{6}$$

There are two very instructive ways of looking at this quantity, which will form a basis for this review. Mathematically, I_S can be written in terms of the Shannon relative entropy. In this sense it represents a distance between the distribution $p(a,b)$ and the product of the marginals $p(a)\times p(b)$. As such, it is intuitively clear that this is a good measure of correlations, since it shows how far a joint distribution is from the product one in which all the correlations have been destroyed, or alternatively, how distinguishable a correlated state is from a completely uncorrelated one. So we have

$$I_S(A:B) = S[p(a,b)||p(a)\times p(b)].$$

Let us now view this from another angle. Suppose that we wish to know the probability of observing b_j if a_i has been observed. This is called a conditional probability and is given by

$$p_{a_i}(b_j) := \frac{p(a_i,b_j)}{p(a_i)}.$$

This motivates us to introduce a conditional entropy, $S_A(B)$, as

$$S_A(B) = -\sum_i p(a_i) \sum_j p_{a_i}(b_j) \ln p_{a_i}(b_j)$$

$$= -\sum_{ij} p(a_i,b_j) \ln p_{a_i}(b_j).$$

This quantity tells us how uncertain we are about the value of B once we have learned about the value of A. Now the Shannon mutual information can be rewritten as

$$I_S(A:B) = S(B) - S_A(B) = S(A) - S_B(A). \tag{7}$$

Hence, the Shannon mutual information, as its name indicates, measures the quantity of information conveyed about the random variable $A(B)$ through measurements of the random variable $B(A)$. This quantity, being posi-

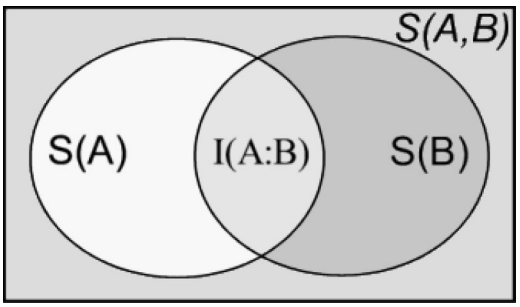

FIG. 2. Venn diagram representation of the joint Shannon entropy of two random variables as well as the marginal Shannon entropies. It is clear that geometrically the Shannon mutual information is obtained by summing the marginal entropies and subtracting the total entropy. It is interesting to note that its generalization fails for three or more random variables.

tive, tells us that the initial uncertainty in $B(A)$ can in no way be increased by making observations on $A(B)$. Note also that, unlike the Shannon relative entropy, the Shannon mutual information is symmetric (see Fig. 2). The following example demonstrates the symmetry of the Shannon mutual information.

Let us briefly go back to our original idea of a surprise to interpret the Shannon mutual information as a measure of correlations. Suppose that one of our friends likes to wear socks of two colors only: red and blue. In addition, we know that her socks are always the same color and that when she gets up in the morning she randomly chooses the color, but we know that she prefers blue to red with the ratio 3:1. So when we meet our friend, before we have looked at the color of her socks, we know that she wears blue socks with the probability $p(b)=0.75$ and red socks with the probability $p(r)=0.25$. However, when we look at one sock and observe, say, the color blue, we immediately know that the other sock must be blue, too. This means that the colors of her two socks are correlated. Before we look at one of the socks, we are uncertain about the color of the other sock by an amount of $-0.75\ln 0.75-0.25\ln 0.25$. But when we look at one of them the uncertainty immediately disappears. We therefore expect that the information we gain about one sock by looking at the color of the other is given by $-0.75\ln 0.75-0.25\ln 0.25$. The Shannon mutual information predicts exactly the same thing. We see that the largest correlations would be if $p(r)=p(b)=0.5$ and this would be $\ln 2$. This, of course, agrees with our intuitive notion of surprise, since before looking at our friend's one sock, we would be completely uncertain about the color of the other sock. By observing its color we obtain the largest possible amount of information (i.e., remove the largest possible uncertainty in this case).

Although it will be seen that the Shannon mutual information is a good measure of correlation between two random variables, its natural generalization to three or more random variables fails. It is easy to see that from three random variables the Shannon mutual information should be of the following form:

$$I_S(A:B:C)=S(A,B,C)-S(A,B)-S(A,C)-S(B,C)+S(A)+S(B)+S(C). \quad (8)$$

However, there exist A,B,C such that $I_S(A:B:C)<0$ (I leave this as an exercise for the reader), and since we regard the amount of correlation as being strictly positive, this is automatically ruled out as a good measure of correlation. A way to sidestep this difficulty is to define mutual information via the relative entropy as $S[p(A,B,C)||p(A)p(B)p(C)]$. This is a positive quantity representing the distance of the joint probability distribution of three random variables A,B,C from the product of the corresponding marginal probability distributions. This, of course, immediately generalizes to any number of random variables. Next I show why the relative entropy and mutual information are also very useful from the dynamical perspective.

C. Classical evolution and relative entropy

The above application of relative entropy to physics via the concept of distinguishability might seem contrived. This is, however, not at all the case, and this section shows the great importance of relative entropy for the dynamics of classical systems. A state of a physical system in classical mechanics can be represented as a vector whose entries are various probabilities for the system to occupy its different possible states. The evolution of this system is seen as the change of these probabilities with time. Hence evolution is a linear transformation of one state into another state, i.e., of a vector into another vector,

$$q_j=\sum_k P(j|k)p_k,$$

where $P(j|k)$ is the conditional probability for the system to change from the state k to the state j. Because the probability has to be conserved ($\Sigma_j q_j=1$), we have $\Sigma_k P(j|k)=1$. Matrices with this simple property, namely, that their entries are positive and columns sum up to 1, are called *stochastic*. The above can be generalized to continuous systems and continuous time evolution, but this will not be relevant for the rest of this review.

A very important property of any measure that aims to quantify the amount of correlation between two random variables (i.e., two states of the same system or two different systems in classical mechanics) is the following: if either or both of the variables undergo a *local* stochastic evolution, then the amount of correlation cannot increase (in fact, it usually decreases). I shall now prove this in the case of the Shannon mutual information, following an approach similar to that given by Everett (1973); see also Penrose's excellent book on statistical mechanics (Penrose, 1973).

First, let us establish without proof two inequalities following from the convex properties of the logarithmic functions (Everett, 1973). Lemma 1 states that entropy

is a concave function, whereas lemma 2 states that relative entropy is a convex function.

Lemma 1. $\Sigma_i P_i x_i \ln \Sigma_i P_i x_i \leqslant \Sigma_i P_i x_i \ln x_i$, where $x_i \geqslant 0$, $P_i \geqslant 0$ and $\Sigma_i P_i = 1$.

Physically, this inequality means that the average uncertainty (negative of the right-hand side) is less than or equal to the uncertainty of the average (negative of the left-hand side); in other words, mixing probability distributions increases entropy. This is a very important property of entropy as a measure of uncertainty, since when we mix probability distributions we expect to increase our uncertainty.

Lemma 2. $\Sigma_i x_i \ln (\Sigma_i x_i / \Sigma_i a_i) \leqslant \Sigma_i x_i \ln (x_i / a_i)$, where $x_i \geqslant 0$ and $a_i \geqslant 0$ for all i.

This is just a statement of the fact that mixing decreases distinguishability. Note that this is in accord with lemma 1, since the more mixed the probability distributions, the less distinguishable they are.

These two simple and self-evident statements lead to a very important result, that the Shannon relative entropy between two probability distributions decreases when the same two undergo a stochastic process. This is a very satisfying property from the physical point of view, where two probability distributions undergoing stochastic changes in fact represent two evolving physical systems. It says that two probability distributions are in some sense closer to each other (i.e., harder to distinguish) after a stochastic process, or analogously, that two physical systems become more alike.

Let us consider a sequence of transition-probability matrices $T_{ij}^n := P_n(i|j)$, where $\Sigma_j T_{ij}^n = 1$ for all n, i, and $0 \leqslant T_{ij}^n \leqslant 1$. We also introduce a sequence of positive measures (i.e., probability distributions) a_i^n having the property that

$$a_j^{n+1} = \sum_i a_i^n T_{ij}^n .$$

Transition probabilities T tell us the probability that at the nth step of evolution the system will "jump" from the jth to the ith state. Thus constructed transition matrices are stochastic for all n. Let us further suppose that we have a sequence of probability distributions p_i^n generated by the action of the above stochastic process, such that

$$p_j^{n+1} = \sum_i p_i^n T_{ij}^n .$$

This is the law describing the system's evolution in time, and the state of the system at time n is given by the probabilities p_i^n. For each of these probability distributions the relative entropy S^n is defined as

$$S^n(p||a) := S(p^n||a^n) = \sum_i p_i^n \ln \frac{p_i^n}{a_i^n} .$$

Let us now prove the following theorem:

Distinguishability never increases,

$$S^{n+1}(p||a) \leqslant S^n(p||a). \tag{9}$$

Proof. Expanding $S^{n+1}(p||a)$ we obtain

$$S^{n+1}(p||a) = \sum_j p_j^{n+1} \ln \frac{p_j^{n+1}}{a_j^{n+1}}$$

$$= \sum_j \left\{ \sum_i p_i^n T_{ij}^n \right\} \ln \frac{\sum_i p_i^n T_{ij}^n}{\sum_i a_i^n T_{ij}^n} .$$

However, using lemma 2 we have the following inequality:

$$\sum_i p_i^n T_{ij}^n \ln \frac{\sum_i p_i^n T_{ij}^n}{\sum_i a_i^n T_{ij}^n} \leqslant \sum_i p_i^n T_{ij}^n \ln \frac{p_i^n T_{ij}^n}{a_i^n T_{ij}^n} .$$

From the above two it follows that

$$S^{n+1}(p||a) \leqslant \sum_j \left\{ \sum_i p_i^n T_{ij}^n \ln \frac{P_i^n}{a_i^n} \right\}$$

$$= \sum_i p_i^n T_{ij}^n \ln \frac{p_i^n}{a_i^n} = \sum_i p_i^n \ln \frac{p_i^n}{a_i^n} = S^n(p||a)$$

and the proof is completed. ■

This property means that a distance between two states cannot increase with time if the states evolve under any stochastic map. The proof can be immediately specialized to the cases in which T is stationary, i.e., T is independent of n, and when T is *doubly stochastic*, i.e., $\Sigma_i T_{ij} = 1$ for all j. A corollary to this important lemma is the following.

Corollary. If we take $p = p(a,b)$ and $a = p(a)p(b)$, and suppose that the stochastic processes acting separately on A and B are uncorrelated, we see that the Shannon mutual information does not increase under these local stochastic processes (by local we mean that they act separately on A and B).

This is a very important, and physically intuitive, property of any measure of correlation; its quantum analog will be of central importance for quantifying quantum correlation between entangled subsystems. This corollary, in fact, can be taken as a guide for a "good" measure of correlation. We can state that any measure of correlation has to be nonincreasing under local stochastic processes. In other words, this means that the only way that systems can become more correlated, i.e., that they gain more information about each other, is if they interact. Without mutual interaction the correlations can only decrease or at best stay the same. The nature of quantum local stochastic processes will form the physical basis for our argument in the next section. A condition similar to property above, but employing quantum stochastic processes, will be a key element in our search for measures of entanglement. When we go to quantum mechanics, the notion of a probability distribution will be replaced by that of a quantum state (i.e., a density matrix), and a stochastic process will become a measurement process in quantum theory. The formulation of probability theory that is most naturally

generalized to quantum states is provided by Kolmogorov (1950), and the quantum generalization expressing similarities with von Neumann's Hilbert-space formulation (von Neumann, 1932) can be found in Mackey (1963; see also Holevo, 1982). However, knowledge of this approach will not be necessary for the rest of the review. Finally it is important to stress that if the local stochastic processes are correlated they virtually become global, and therefore the correlations between the systems can increase as well as decrease.

D. Schmidt decomposition and quantum dynamics

The difference between classical and quantum physics can be seen in the fact that quantum states are described by a density matrix ρ (and not just vectors). The density matrix is a positive semidefinite Hermitian matrix, whose trace is unity (representing the fact that all the probabilities add up to 1). An important class of density matrices is the idempotent one, i.e., $\rho^2=\rho$. The states these matrices represent are called pure states. When there is no uncertainty in the knowledge of the system's state its state is then pure. Another important notion is that of a composite system. A composite quantum system is one that consists of a number of quantum subsystems. When those subsystems are entangled it is impossible to ascribe a definite state vector to any one of them. The most often cited entangled system is a pair of two photons, being in the Einstein-Podolsky-Rosen state (Einstein *et al.*, 1935; Bell, 1987). The composite system is then mathematically described by

$$|\Psi\rangle=\frac{1}{\sqrt{2}}(|\uparrow\rangle|\downarrow\rangle+|\downarrow\rangle|\uparrow\rangle), \tag{10}$$

where the first ket in either product belongs to one photon and the second to the other. The property that is described is the direction of spin or polarization along the z axis, which can either be up ($|\uparrow\rangle$) or down ($|\downarrow\rangle$). A two-level system of this type is the quantum analog of a bit, which we shall henceforth call a *qubit*. We can immediately see that neither of the photons possesses a definite state vector. The best that one can say is that if a measurement is made on one photon and it is found to be in the up state, for example, then the other photon is certain to be in the down state. This idea cannot be applied to a general composite system unless the former is written in a special form. This motivates us to introduce the so-called Schmidt decomposition (Schmidt, 1907), which not only is mathematically convenient, but also gives a deeper insight into correlations between the two subsystems.[1]

According to the rules of quantum mechanics the state vector of a composite system, consisting of subsystems U and V, is represented by a vector belonging to the tensor product of the two Hilbert spaces $\mathcal{H}_U\otimes\mathcal{H}_V$. The general state of this system can be written as a linear superposition of products of individual states:

$$|\Psi\rangle=\sum_n\sum_m c_{nm}|u_n\rangle|v_m\rangle, \tag{11}$$

where $\{|u_n\rangle\}_{n=1}^N$ and $\{|v_m\rangle\}_{m=1}^N$ are the orthonormal basis of the subsystems U and V, respectively, whose dimensions are $\dim U=N$ and $\dim V=M$. This state can always be written in the so-called Schmidt form:

$$|\Psi\rangle=\sum_n g_n|u_n'\rangle|v_n'\rangle, \tag{12}$$

where $|u_n'\rangle$ and $|v_n'\rangle$ are orthonormal bases for U and V, respectively. Note that in this form the correlations between the two subsystems are fully displayed. If U is found in the state $|u_2'\rangle$, for example, then the state of V is $|v_2'\rangle$. This is clearly a multistate generalization of the Einstein-Podolsky-Rosen state mentioned earlier.

I shall now prove this assertion by showing how to derive Eq. (12) from Eq. (11). To that end, let us assume that $M>N$, which in no way affects our line of argument since the procedure is symmetric with respect to the subsystems. Then we have the following five steps.

(1) First we construct a density matrix describing $|\Psi\rangle=\Sigma_n\Sigma_m c_{nm}|u_n\rangle|v_m\rangle$. Once the density matrix is known, all the properties of the system can be deduced from it. Moreover, ensembles that are prepared differently but have the same density matrix are statistically indistinguishable and therefore equivalent. Generally, if we have a mixed state involving vectors $|\Psi_1\rangle,|\Psi_2\rangle,\ldots,|\Psi_D\rangle$ with corresponding classical probabilities $w_1,w_2,\ldots,w_3$, then the density matrix is defined to be
$$\rho=\sum_{d=1}^{D} w_d|\Psi_d\rangle\langle\Psi_d|.$$
Since in our case $|\Psi\rangle$ is a pure state, the density matrix is a projection operator onto $|\Psi\rangle$, i.e.,
$$\rho=|\Psi\rangle\langle\Psi|=\sum_{nm}\sum_{pq}\rho_{nmpq}|u_n\rangle\langle u_p|\otimes|v_m\rangle\langle v_q|,$$
where $\rho_{nmpq}=c_{nm}c_{pq}^*$. If, however, we wish to deal with only one of the subsystems, then we employ the concept of the reduced density matrix.

(2) We find the reduced density matrix of the subsystem U, obtained by tracing ρ over all states of the subsystem V, so that
$$\rho_U=\sum_q\langle v_q|\rho|v_q\rangle=\sum_{nm}\sum_p\rho_{nmpm}|u_n\rangle\langle u_p|.$$
Note that the partial trace (or the trace itself) does not depend on the choice of basis. Partial tracing is analogous to finding marginal probability distributions from a joint probability distribution in classical probability theory. The crucial step in the Schmidt decomposition is diagonalizing the above. I shall call the eigenvalues of ρ_U $|g_1|^2,|g_2|^2,\ldots,|g_N|^2$, and the corresponding eigenvectors $|u_1'\rangle,|u_2'\rangle,\ldots,|u_N'\rangle$.

[1]In the context of quantum theory see Everett (1957; 1973, p. 3). A graduate-level textbook by Peres (1993, Chap. 5) includes a brief description of the Schmidt decomposition.

(3) We then reexpress the above in terms of $|u'\rangle$'s, i.e.,

$$|\Psi\rangle=\sum_n \sum_m c'_{nm}|u'_n\rangle|v_m\rangle.$$

(4) Now we construct a new orthonormal basis of the subsystem V such that each new vector is a "clever" linear superposition of the old ones, so that

$$|v'_i\rangle=\sum_m \frac{c'_{im}}{g_i}|v_m\rangle.$$

The matrix given by the coefficients c'_{im}/g_i is unitary, which is why the new basis is orthonormal.

(5) The Schmidt decomposition of $|\Psi\rangle$ is now given by

$$|\Psi\rangle=\sum_n g_n|u'_n\rangle|v'_n\rangle.$$

There are two important observations to be made, which are fundamental to understanding the correlation between two subsystems in a joint pure state.

- The reduced density matrices of both subsystems, written in the Schmidt basis, are diagonal and have the same positive spectrum. In particular, the overall density matrix is given by

$$\rho=\sum_{nm} g_n g_m^* |u'_n\rangle\langle u'_m|\otimes|v'_n\rangle\langle v'_m|$$

whereas the reduced ones are

$$\rho_U=\sum_m \langle v'_m|\rho|v'_m\rangle=\sum_n |g_n|^2|u'_n\rangle\langle u'_n|,$$

$$\rho_V=\sum_n \langle u'_n|\rho|u'_n\rangle=\sum_m |g_m|^2|v'_m\rangle\langle v'_m|.$$

- If a subsystem is N dimensional it can then be entangled with no more than N orthogonal states of another one.

I should like to point out that the Schmidt decomposition is, in general, impossible for more than two entangled subsystems. To clarify this, consider three entangled subsystems as an example. Here our intention would be to write a general state such that by observing the state of one of the subsystems we could instantaneously and with certainty know the state of the other two. But this is impossible in general, for the presence of the third system makes the prediction uncertain. Loosely speaking, while we know the state of one of the subsystems, the other two might still be entangled and cannot have definite vectors associated with them [an exception to this general rule is, for example, a state of the Greenberger-Horne-Zeilinger type $(1/\sqrt{2})(|\uparrow\rangle|\uparrow\rangle|\uparrow\rangle+|\downarrow\rangle|\downarrow\rangle|\downarrow\rangle)$]. Clearly, involvement of even more subsystems complicates this analysis even further and produces, so to speak, an even greater mixture and uncertainty. The same reasoning applies to mixed states of two or more subsystems (i.e., states whose density operator is not idempotent $\rho^2\neq\rho$), for which we cannot have the Schmidt decomposition in general. This reason alone is responsible for the fact that the entanglement of two subsystems in a pure state is simple to understand and quantify, while for mixed states, or states consisting of more than two subsystems, the question is much more involved.

Let us now consider the way in which quantum systems evolve. An isolated system, of course, follows a unitary dynamics generated by Schrödinger's equation (nonrelativistic). This evolution is fully reversible (manifesting itself in the fact that the quantum entropy does not increase during this process, as we shall see below). However, we know that most of the processes in Nature are irreversible (think of spontaneous emission and the nonexistence of its reverse, "spontaneous absorption"). These processes are nonunitary and arise from the interaction of the system with the environment; thus the system is no longer closed. Mathematically, the evolution of a quantum state is then most generally of the form (Davies, 1976)

$$\rho'=\sum_\alpha A_\alpha \rho A_\alpha^\dagger, \tag{13}$$

where, because of the conservation of probability, or, more precisely, trace preservation, $\Sigma_\alpha A_\alpha^\dagger A_\alpha=1$. The above map is the most general completely positive (trace-preserving) linear map (CP map; Choi, 1975). Positivity means that density matrices are mapped into density matrices (strictly speaking, positive operators are mapped onto positive operators). To define "complete," we first need to introduce the idea of an extended state. By extension of a state I mean any state on a larger Hilbert space that reduces itself to the original state when the extended part is traced out. In turn, completeness means that any extension of the density matrix is also mapped into a density matrix. To clarify this I shall present a few examples of CP maps.

- Projectors are Hermitian idempotent operators ($P^\dagger=P$ and $P^2=P$) and the evolution of the form $\rho\rightarrow\Sigma_i P_i\rho P_i$ is a CP map;
- Addition of another system to ρ is also a CP map, $\rho\rightarrow\rho\otimes\rho_1$;
- Let $E_i\geqslant 0$ and $\Sigma_i E_i=I$. Then $\rho\rightarrow p_k:=\mathrm{Tr}(\rho E_k)$ is a CP map that generates a probability distribution from a density matrix.
- Unitary evolution is a special case of a CP map, where only one operator is present in the sum, i.e., $U\rho U^\dagger$.

I leave it for the reader to show that the above CP maps can indeed be written in the form of Eq. (13). We shall see other examples in the next subsection.

Remarkably not all positive maps are completely positive, transposition being a well-known example. Positivity of transposition follows from the fact that for any state ρ, its transposition $\rho^T\geqslant 0$. However, a counterexample to completeness comes from, for example, a singlet state of two subsystems A and B. Namely, if we transpose only A (or B), then the resulting operator is not positive (so that it is not a physical state), i.e., $\rho_{AB}^{T_A}<0$. Confirmation of this is left as an exercise.

The reader might wonder what the physical implementation of the canonical form $\Sigma_\alpha A_\alpha\rho A_\alpha^\dagger$ is. I shall

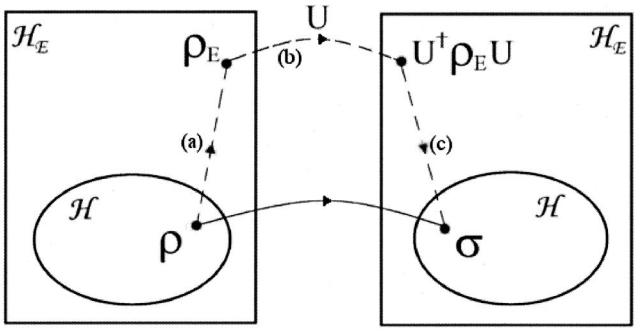

FIG. 3. Completely positive trace-preserving map. The most general evolution in quantum mechanics is represented by a completely positive trace-preserving map (CP map). This figure shows two equivalent forms for such a map: (a) the canonical form $A(\cdot)A^\dagger$; (b) the extension to a larger Hilbert space $\mathcal{H}_E$ and an appropriate unitary transformation therein. The connection is explained in the text.

now introduce another kind of CP map that will explain its physical importance and will be crucial for the rest of the review. Loosely stated, any CP map can be represented as a unitary transformation on a higher Hilbert space (see Fig. 3). That is, from Schmidt decomposition we know that a density matrix can be represented as a "reduction" of a state in an enlarged Hilbert space. Suppose that $\rho\in\mathcal{H}$ and that $\rho_E\in\mathcal{H}\otimes\mathcal{H}_a$ is an "extension" of the state ρ such that $\mathrm{Tr}_a\rho_E=\rho$. Then a CP map $\sigma=\Phi(\rho)$ can be represented as

$$\rho\rightarrow\rho_E\rightarrow U\rho_E U^\dagger\rightarrow\mathrm{Tr}_a(U\rho_E U^\dagger)=\sigma. \tag{14}$$

Here we have first "lifted" ρ to ρ_E, then evolved ρ_E unitarily into $U\rho_E U^\dagger$, which, after tracing over the Hilbert-space extension (i.e., lowering), yields the final state σ as in Fig. 3. The fact that for any CP map there exists a unitary operator U that will execute this map on some higher Hilbert space is guaranteed by a theorem proved independently by Kraus (1983) and Ozawa (1984; see Schumacher, 1996 for a modern presentation). I shall now present only a plausibility argument for this correspondence. Let $\rho_E=\rho\otimes|0\rangle\langle 0|_a$ where $|0\rangle\langle 0|_a\in\mathcal{H}_a$. Then

$$\sigma=\mathrm{Tr}_a(U\rho\otimes|0\rangle\langle 0|_a U^\dagger)=\sum_i\langle i|_a U\rho\otimes|0\rangle\langle 0|_a U^\dagger|i\rangle_a$$

$$=\sum_i\langle i|U|0\rangle\rho\langle 0|U^\dagger|i\rangle,$$

which has the same form as Eq. (13) providing that we define $A_i:=\langle i|U|0\rangle$. Thus, given a unitary evolution on the extended Hilbert space, we can always find corresponding positive operators that describe the evolution of the original system. Note that the choice of the operators is not unique.

Finally, I should like to discuss another frequently used concept that is in some sense derived from the notion of the CP map. It can be loosely stated that the CP map represents the evolution of a quantum system when we do not update the knowledge of its state based on the particular measurement outcome. This is why we have a summation over all measurements in Eq. (13). If, on the other hand, we know that the outcome corresponding to the operator $A_j^\dagger A_j$ occurs, then the state of the system immediately afterwards is given by $A_j\rho A_j^\dagger/\mathrm{tr}(A_j^\dagger A_j\rho)$. This type of measurement is the most general one and is commonly referred to as the positive operator valued measure (POVM). It is positive because operators of the form $A^\dagger A$ are always positive for any operator A and taking the trace of it together with any density matrix generates a positive number (i.e., a probability for that particular measurement outcome). For a more detailed overview of POVM's see Peres (1993). The concept of the POVM will play a significant role when defining the quantum relative entropy in the next subsection.

E. Quantum relative entropy

When two subsystems become entangled, the composite state can be expressed as a superposition of the products of the corresponding Schmidt basis vectors. From Eq. (12) it follows that the ith vector of either subsystem has a probability of $|g_i|^2$ associated with it. We are, therefore, uncertain about the state of each subsystem, the uncertainty being larger if the probabilities are evenly distributed. Since the uncertainty in the probability distribution is naturally described by the Shannon entropy, this classical measure can also be applied in quantum theory. In an entangled system this entropy is related to a single observable. The general state of a quantum system, as I have already remarked, is described by its density matrix ρ. If A is an observable pertaining to the system described by ρ, then by the spectral decomposition theorem $A=\Sigma_i a_i P_i$, where P_i is the projection onto the state with the eigenvalue a_i. The probability of obtaining the eigenvalue a_j is given by $p_j=\mathrm{Tr}(\rho P_j)=\mathrm{Tr}(P_j\rho)$. The uncertainty in a given observable can now be expressed through the Shannon entropy. Let the observables A and B, pertaining to the subsystems U and V, respectively, have a discrete, nondegenerate spectrum, with corresponding probabilities $p(a_i)$ and $p(b_j)$ of observables A being a_i and B being b_j. In addition, let the joint probability be $p(a_i,b_j)$. Then

$$S(A)=-\sum_i p(a_i)\ln p(a_i) \tag{15}$$

$$=-\sum_{ij} p(a_i,b_j)\ln\sum_j p(a_i,b_j), \tag{16}$$

$$S(B)=-\sum_j p(b_j)\ln p(b_j) \tag{17}$$

$$=-\sum_{ij} p(a_i,b_j)\ln\sum_i p(a_i,b_j), \tag{18}$$

$$S(A,B)=-\sum_{ij} p(a_i,b_j)\ln p(a_i,b_j), \tag{19}$$

where I have used the fact that $\Sigma_j p(a_i,b_j)=p(a_i)$ and $\Sigma_i p(a_i,b_j)=p(b_j)$. We have seen that a signature of correlation is that the sum of the uncertainties in the individual subsystems is greater than the uncertainty in the total state. Hence, the Shannon mutual information is a good indicator of how much the two given observables are correlated. However, this quantity, as it is inherently classical, describes the correlations between single observables only. The quantity that is related to the correlations in the overall state as a whole is the von Neumann mutual information. Since it is assigned to the state as a whole, it is of little surprise that it depends on the density matrix. First, however, I define the von Neumann entropy (von Neumann, 1932), which can be considered as the proper quantum analog of the Shannon entropy (Wehrl, 1978; Ohya and Petz, 1993; Ingarden *et al.*, 1997).

Definition. The *von Neumann entropy* of a quantum system described by a density matrix ρ is defined as

$$S_N(\rho) := -\mathrm{Tr}(\rho \ln \rho) \tag{20}$$

(I shall drop the subscript N whenever there is no possibility of confusion). The Shannon entropy is equal to the von Neumann entropy only when it describes the uncertainties in the values of the observables that commute with the density matrix, i.e., the Schmidt observables. Otherwise

$$S(A) \geqslant S_N(\rho),$$

where A is any observable of a system described by ρ. This means that there is more uncertainty in a single observable than in the whole of the state, a fact that entirely contradicts our expectations.

I now discuss a relation concerning the entropies of two subsystems. One part of it is somewhat analogous to its classical counterpart, but instead of referring to observables it is related to the two states. This inequality is called the Araki-Lieb inequality (Araki and Lieb, 1970) and is one of the most important results in the quantum theory of correlations. Let ρ_A and ρ_B be the reduced density matrices of subsystems A and B, respectively, and let ρ be the matrix of a composite system; then

$$S_N(\rho_A)+S_N(\rho_B) \geqslant S_N(\rho) \geqslant |S_N(\rho_A)-S_N(\rho_B)|.$$

Physically, the left-hand side implies that we have more information (less uncertainty) in an entangled state than if the two states are treated separately. This arises naturally, since by treating the subsystems separately we have neglected the correlations (entanglement). We note that if the composite system is in a pure state, then $S(\rho)=0$, and from the right-hand side it follows that $S(\rho_A)=S(\rho_B)$ [cf. Schmidt decomposition Eq. (12)]. To appreciate the extent to which this is a counterintuitive result, consider the following example. Suppose a two-level atom is interacting with a single mode of an electromagnetic field as in the Jaynes-Cummings model (Jaynes and Cummings, 1963). If the overall state is initially pure, and the whole system is isolated, then the entropies of the atom and the field are equally uncertain at all times. This is not expected, since the atom has only two degrees of freedom and the field infinitely many. However, it is so, as, by the second observation, the atom, as a two-dimensional subsystem, is entangled with only two dimensions of the field.

I present without proofs two important properties of entropy that will be used in later sections (Wehrl, 1978). These are

$$\text{(1) additivity:} \quad S_N(\rho_A \otimes \rho_B) = S_N(\rho_A)+S_N(\rho_B); \tag{21}$$

$$\text{(2) concavity:} \quad S_N\left(\sum_i \lambda_i \rho_i\right) \geqslant \sum_i \lambda_i S_N(\rho_i). \tag{22}$$

The first property is the same as in classical information theory, namely, the entropies of independent systems add up. The concavity simply reflects the fact that mixing increases uncertainty.

Following the definition of the Shannon mutual information, I introduce the von Neumann mutual information, which refers to the correlation between whole subsystems rather than that relating only two observables.

Definition. The *von Neumann mutual information* between two subsystems ρ_U and ρ_V of a joint state ρ_{UV} is defined as

$$I_N(\rho_U:\rho_V;\rho_{UV}) = S_N(\rho_U)+S_N(\rho_V)-S_N(\rho_{UV}). \tag{23}$$

As in the case of the Shannon mutual information this quantity can be interpreted as a distance between two quantum states. For this we first need to define the von Neumann relative entropy, in a direct analogy with the Shannon relative entropy [in fact, this quantity was first considered by Umegaki (1962), but for consistency reasons I name it after von Neumann; I shall also refer to it as the quantum relative entropy].

Definition. The *von Neumann relative entropy* between the two states σ and ρ is defined as

$$S_N(\sigma\|\rho) = \mathrm{Tr}\sigma(\ln\sigma - \ln\rho). \tag{24}$$

This measure also has the same statistical interpretation as its classical analog: it tells us how difficult it is to distinguish the state σ from the state ρ (Hiai and Petz, 1991). To that end, suppose we have two states σ and ρ. How can we distinguish them? We can choose a POVM $\Sigma_{i=1}^{M} A_i = \mathbf{1}$ that generates two distributions via

$$p_i = \mathrm{tr}A_i\sigma, \tag{25}$$

$$q_i = \mathrm{tr}A_i\rho, \tag{26}$$

and use classical reasoning to distinguish these two distributions. However, the choice of POVM is not unique. It is therefore best to choose that POVM which distinguishes the distributions most, i.e., for which the *classical* relative entropy is largest. Thus we arrive at the following quantity:

$$S_1(\sigma\|\rho) := \sup_{A\text{'s}}\left\{\sum_i \mathrm{tr}A_i\sigma \ln \mathrm{tr}A_i\sigma - \mathrm{tr}A_i\sigma \ln \mathrm{tr}A_i\rho\right\},$$

where the supremum is taken over all POVM's. The above is not the most general measurement that we can make, however. In general we have N copies of σ and ρ in the state

$$\sigma^N = \underbrace{\sigma\otimes\sigma\otimes\cdots\otimes\sigma}_{\text{total of } N \text{ terms}}, \tag{27}$$

$$\rho^N = \underbrace{\rho\otimes\rho\otimes\cdots\otimes\rho}_{\text{total of } N \text{ terms}}. \tag{28}$$

We may now apply a POVM $\Sigma_i A_i = \mathbf{1}$ acting on σ^N and ρ^N. Consequently we define a new type of relative entropy,

$$S_N(\sigma\|\rho) := \sup_{A\text{'s}}\left\{\frac{1}{N}\sum_i \mathrm{tr}A_i\sigma^N \ln \mathrm{tr}A_i\sigma^N - \mathrm{tr}A_i\sigma^N \ln \mathrm{tr}A_i\rho^N\right\}. \tag{29}$$

Now it can be shown that (Donald, 1986, 1987)

$$S(\sigma\|\rho) \geqslant S_N, \tag{30}$$

where $S(\sigma\|\rho)$ is the quantum relative entropy. (This really is a consequence of the fact that the relative entropy does not increase under general CP maps, a fact that will be proven later in this subsection.) Equality is achieved in Eq. (30) if and only if σ and ρ commute (Fuchs, 1996). However, for any σ and ρ it is true that (Hiai and Petz, 1991)

$$S(\sigma\|\rho) = \lim_{N\to\infty} S_N.$$

In fact, this limit can be achieved by projective measurements that are independent of σ (Hayashi, 2001). From these considerations it would naturally follow that the probability of confusing two quantum states σ and ρ (after performing N measurements on ρ) is (for large N)

$$P_N(\rho\to\sigma) = e^{-NS(\sigma\|\rho)}. \tag{31}$$

We should like to stress here that classical statistical reasoning applied to distinguishing quantum states leads to the above formula. There are, however, other approaches. Some take Eq. (31) for their starting point and then derive the rest of the formalism thenceforth (Hiai and Petz, 1991). Others assume a set of axioms that must be satisfied by the quantum analog of the relative entropy (for example, it should reduce to the classical relative entropy if the density operators commute, i.e., if they are classical) and then derive Eq. (31) as a consequence (Donald, 1986, 1987). In any case, as I have argued here, there is strong reason to believe that the quantum relative entropy $S(\sigma\|\rho)$ plays the same role in quantum statistics as the classical relative entropy plays in classical statistics (see also the review by Schumacher and Westmoreland, 2000).

The von Neumann mutual information can now be understood as the distance of the state ρ_{UV} from the uncorrelated state $\rho_U\otimes\rho_V$,

$$I_N(\rho_U:\rho_V;\rho_{UV}) = S_N(\rho_{UV}\|\rho_U\otimes\rho_V).$$

The quantum relative entropy will be the most important quantity in classifying and quantifying quantum correlations. It will be seen that this quantity does not increase under CP maps, which are quantum analogs of the stochastic processes. I list three properties of the relative entropy without proof.

(F1) Unitary operations leave $S(\sigma\|\rho)$ invariant, i.e., $S(\sigma\|\rho) = S(U\sigma U^\dagger\|U\rho U^\dagger)$. Unitary transformations represent a change of basis (i.e., a change in our "perspective") and the distance between two states should not (and does not in this case) change under this.

(F2) $S(\mathrm{Tr}_p\sigma\|\mathrm{Tr}_p\rho) \leqslant S(\sigma\|\rho)$, where Tr_p is a partial trace. Tracing over a part of the system leads to a loss of information. The less information we have about two states, the harder they are to distinguish, which is what this inequality says. (This property is closely related to the strong subadditivity of relative entropy as shown in Lieb and Ruskai, 1973; see also a more recent proof by Lesniewski and Ruskai, 1999.)

(F3) The relative entropy is additive $S(\sigma_1\otimes\sigma_2\|\rho_1\otimes\rho_2) = S(\sigma_1\|\rho_1) + S(\sigma_2\|\rho_2)$. This inequality is a consequence of the additivity of entropy itself.

These properties of relative entropy have profound implications for the evolution of quantum systems, as I now show.

Quantum distinguishability never increases. For any completely positive, trace-preserving map Φ, given by $\Phi\sigma = \Sigma V_i\sigma V_i^\dagger$ and $\Sigma V_i^\dagger V_i = 1$, we have $S(\Phi\sigma\|\Phi\rho) \leqslant S(\sigma\|\rho)$.

I shall first present a physical argument as to why we should expect this theorem to hold. As I have discussed, a CP map can be represented as a unitary transformation on an extended Hilbert space. According to (F1), unitary transformations do not change the relative entropy between two states. However, after this, we have to perform a partial tracing to go back to the original Hilbert space, which, according to (F2), decreases the relative entropy as some information is invariably lost during this operation. Hence the relative entropy decreases under any CP map. I now formalize this proof.

Proof. I have discussed the fact that a CP map can always be represented as a unitary operation + partial tracing on an extended Hilbert space $\mathcal{H}\otimes\mathcal{H}_n$, where $\dim\mathcal{H}_n = n$ (Lindblad, 1974, 1975). Let $\{|i\rangle\}$ be an orthonormal basis in $\mathcal{H}_n$ and $|\alpha\rangle$ be a unit vector. I define

$$W = \sum_i V_i\otimes|i\rangle\langle\alpha|. \tag{32}$$

Then $W^\dagger W = 1\otimes P_\alpha$, where $P_\alpha = |\alpha\rangle\langle\alpha|$, and there is a unitary operator U in $\mathcal{H}\otimes\mathcal{H}_n$ such that $W = U(1\otimes P_\alpha)$ (Reed and Simon, 1980). Consequently

$$U(A\otimes P_\alpha)U^\dagger = \sum_{ij} V_i A V_j^\dagger\otimes|i\rangle\langle j|, \tag{33}$$

so that

$$\mathrm{Tr}_2\{U(A\otimes P_\alpha)U^\dagger\} = \sum_i V_i A V_i^\dagger.$$

This shows that the unitary and $\Sigma_i V_i\rho V_i^\dagger$ representations are equivalent. Now using properties (F2), then (F1), and finally (F3) we find

$$S[\mathrm{Tr}_2\{U(\sigma\otimes P_\alpha)U^\dagger\}\|\mathrm{Tr}_2\{U(\rho\otimes P_\alpha)U^\dagger\}]$$
$$\leqslant S[U(\sigma\otimes P_\alpha)U^\dagger\|U(\rho\otimes P_\alpha)U^\dagger]$$
$$= S(\sigma\otimes P_\alpha\|\rho\otimes P_\alpha)=S(\sigma\|\rho). \quad (34)$$

This proves the result. ■

Corollary. Since for a complete set of orthonormal projectors P, $\Sigma_i P_i\sigma P_i$ is a CP map, then

$$\sum_i S(P_i\sigma P_i\|P_i\rho P_i)\leqslant S(\sigma\|\rho). \quad (35)$$

[The sum can be taken outside as it can be easily shown that $S(\Sigma_i P_i\sigma P_i\|\Sigma_i P_i\rho P_i)=\Sigma_i S(P_i\sigma P_i\|P_i\rho P_i)$.] Now from (F1), (F2), (F3), and Eq. (35) we have the following theorem.

Theorem 5. If $\sigma_i=V_i\sigma V_i^\dagger$ then $\Sigma S(\sigma_i\|\rho_i)\leqslant S(\sigma\|\rho)$, where $\rho_i=V_i\rho V_i^\dagger/\mathrm{tr}(V_i\rho V_i^\dagger)$.

Proof. Equations (32) and (33) are introduced as in the previous proof. From Eq. (33) we have

$$\mathrm{Tr}_2\{1\otimes P_i U(A\otimes P_\alpha)U^\dagger 1\otimes P_i\}=V_i A V_i^\dagger,$$

where $P_i=|i\rangle\langle i|$. Now, from (F2), the Corollary, and (F3) it follows that

$$\sum_i S[\mathrm{Tr}_2\{1\otimes P_i U(\sigma\otimes P_\alpha)U^\dagger 1\otimes P_i\}\|\mathrm{Tr}_2\{1\otimes P_i U(\rho\otimes P_\alpha)U^\dagger 1\otimes P_i\}]$$
$$\leqslant\sum_i S[1\otimes P_i U(\sigma\otimes P_\alpha)U^\dagger 1\otimes P_i\|1\otimes P_i U(\rho\otimes P_\alpha)U^\dagger 1\otimes P_i]$$
$$\leqslant S[U(\sigma\otimes P_\alpha)U^\dagger\|U(\rho\otimes P_\alpha)U^\dagger]=S(\sigma\otimes P_\alpha\|\rho\otimes P_\alpha)=S(\sigma\|\rho). \quad (36)$$

This proves theorem 5. ■

This theorem will be important in the next section. A simple consequence of the fact that the quantum relative entropy itself does not increase under CP maps is that correlations (as measured by the quantum mutual information) also cannot increase, but this is now true under local CP maps.

Correlations cannot increase without interaction. Correlations, as measured by the von Neumann mutual information, do not increase during local complete measurements carried out on two entangled quantum systems.

The Shannon mutual information, although having this desired property, does not distinguish between quantum and classical correlations (rather, it measures total correlations). In order to distinguish between quantum and classical, we shall have to introduce the possibility of classical communication between A and B. This will allow classical correlations to increase while leaving quantum correlations intact, as will be seen in the following section. Now let us put the theory developed so far to practical use in communication.

Digression on the second law of thermodynamics. The second law of thermodynamics states that the entropy of an isolated system never decreases. This does not follow directly from the inability of the quantum relative entropy to increase under CP maps. Strictly speaking, an isolated system in quantum mechanics evolves unitarily and therefore its entropy never changes. Under CP maps, however, the entropy can both increase and decrease. If, however, the state ρ is maximally mixed—I/n for example—then the quantum relative entropy is given by

$$S(\sigma\|\rho)=\ln n-S(\sigma). \quad (37)$$

If in addition the evolution is such that I/n is the equilibrium state, then the monotonic decrease in the quantum relative entropy implies a monotonic increase in $S(\sigma)$, just as in the second law of thermodynamics. Otherwise the entropy itself could both increase and decrease. A detailed discussion of the statistical foundations of the second law can be found in Tollman's classic work, *The Principles of Statistical Mechanics* (Tolman, 1938).

III. QUANTUM COMMUNICATION: CLASSICAL USE

The central objective of communication theory is to allow a person, often referred to as Alice, to communicate accurately with another person, called Bob, even in the presence of noise. Alice encodes her message into a number of different (distinguishable) states, with each state representing a different symbol in the message. For example, Alice encodes the bit value 1 into the excited state of a two-level atom and sends this atom to Bob. On its way to Bob the atom may transform into its ground state due to either stimulated or spontaneous emission, thereby giving Bob the impression that Alice transmitted 0. This unwanted state transition is a form of channel noise.

The key question is: what is the largest amount of information (per symbol) that Alice can send to Bob, i.e., what is the capacity of the communication channel taking into account any possible noise? In classical information theory the capacity for communication is given by the mutual information between Alice's sent message and Bob's received message (Shannon and Weaver, 1949). This is intuitively clear, since mutual information quantifies correlations between sent and received messages and it thus tells us how faithful the transmission is.

If we use quantum states to encode symbols, then the capacity is not given by the quantum mutual information introduced earlier. We derive a new quantity for this purpose called the Holevo bound (Holevo, 1973; for the continuous case see Yuen and Ozawa, 1993). The benefit of performing the full quantum derivation is that this is a more fundamental approach to information processing. We can then deduce the classical capacity as a special case.

A. The Holevo bound

A quantum communication channel consists of a number N of quantum systems prepared in states $\rho_1,\rho_2,\ldots,\rho_N$ and whatever physical medium is used to send the states from Alice to Bob. These states encode N different symbols with certain *a priori* probabilities, $p_1,p_2,\ldots,p_N$. Bob then performs a set of measurements to determine the correct sequence of states comprising Alice's symbols, which he can then use to reconstruct the entire message (Ingarden, 1976). If the states suffer no error on the way to Bob, then the channel is called noiseless; otherwise it is called noisy. I consider only the capacity of a noiseless quantum communication channel, since the generalization to a noisy channel is straightforward.

Let $S(\rho)=-\mathrm{Tr}\rho\ln\rho$ be the standard von Neumann entropy of a density matrix ρ. The capacity of a quantum communication channel is then defined as

$$\mathbf{C}:=\max_{\{p\}} C(\{p\},\rho),$$

where

$$C(\{p\},\rho)=S\left(\sum_i p_i\rho_i\right)-\sum_i p_i S(\rho_i) \tag{38}$$

is the *Holevo bound*. Note that the above can be expressed succinctly as

$$C(\{p\},\rho)=\sum_i p_i S(\rho_i\|\rho), \tag{39}$$

where $S(\|)$ is the von Neumann relative entropy and $\rho=\Sigma_i p_i\rho_i$. When there is no possibility of confusion I write $C(\{p\},\rho)\equiv C(\{p\})$. The reader may ask why we need to maximize symbol probabilities in order to compute the capacity. This is because the channel can be used with different input probabilities and the capacity represents the maximum that can be communicated using this channel.

To see the physical motivation behind this quantity consider N states $\rho_1,\ldots,\rho_N$ sent by Alice to Bob according to probabilities $p_1,\ldots,p_N$, respectively. Bob now performs a set of complete measurements POVM $\Sigma_i E_i=I$, where $E_i\geqslant 0$, in order to determine which state was sent to him (a complete measurement is like a CP map, but one in which we record each of the outcomes). The *accessible information* to Bob is given by the mutual information between his measurement and $\rho_1,\ldots,\rho_N$ (Holevo, 1973; Davies, 1976). This quantity tells us how well Bob's measurement can distinguish between the message states and is given by

$$I(E:\rho)=\left\{\sum_i -\mathrm{Tr}(\rho E_i)\ln[\mathrm{Tr}(\rho E_i)] + \sum_j p_j\mathrm{Tr}(\rho_j E_i)\ln[\mathrm{Tr}(\rho_j E_i)]\right\}.$$

The rationale behind this expression is that the uncertainty in the message before any measurement is performed is given by the first term, and the second term represents the uncertainty after the measurement has identified (partially in general) the message states. The Holevo bound is an upper bound to the above accessible information, i.e.,

$$S\left(\sum_i p_i\rho_i\right)-\sum_i p_i S(\rho_i)\geqslant\max_E I(E:\rho). \tag{40}$$

This equality is saturated if and only if $[\rho_i,\rho_j]=0$ for all i and j. Therefore, since the Holevo bound is an upper bound to accessible information that Bob can gain about Alice's message, we identify its maximum over all possible initial probabilities with the classical capacity of a quantum channel.

The Holevo bound has an even more suggestive form: the uncertainty in the initial message is $S(\rho)$, but after the states are correctly identified the average uncertainty is $\Sigma_i p_i S(\rho_i)$. The difference between these two quantities when maximized over all $p_i s$ is the classical communication capacity of a quantum channel. Note that one of the most profound implications of the Holevo bound is that a quantum bit cannot store more information than a classical bit. In spite of this limitation, quantum information processing is more efficient than its classical analog. This is due to the different nature of information encoding, which is reflected in the existence of superpositions of different states as well as entanglement between different qubits (see also the section on dense coding).

Proof of the Holevo bound in Eq. (40). The Holevo bound is a direct consequence of the fact that the quantum relative entropy does not increase under CP maps as in theorem 1. [Note that Holevo's original proof is much more complicated and does not involve using the quantum relative entropy. Here I follow Yuen and Ozawa in spirit (1993); for an alternative proof see King and Ruskai, 2001.] One such map is

$$\tau(A)=\frac{1}{n}\mathrm{Tr}(A),$$

where A is any $n\times n$ positive matrix. This leads to the Peierls-Bogoliubov inequality (Bhatia, 1997);

$$\tau(A)[\ln\tau(A)-\ln\tau(B)]\leqslant\tau(A\ln A-A\ln B). \tag{41}$$

To prove the Holevo bound I first use that fact that (theorem 5)

$$S(\rho_i\|\rho)\geqslant\sum_j S(A_j\rho_i A_j^\dagger\|A_j\rho A_j^\dagger).$$

The Peierls-Bogoliubov inequality now implies that

$$S(A_j\rho_i A_j^\dagger \| A_j\rho A_j^\dagger) \geq \mathrm{Tr}(A_j\rho_i A_j^\dagger)\{\ln[\mathrm{Tr}(A_j\rho_i A_j^\dagger)] - \ln[\mathrm{Tr}(A_j\rho A_j^\dagger)]\} = p(j|i)[\ln p(j|i) - \ln p(j)],$$

where $p(j|i)=\mathrm{Tr}\{A_j\rho_i A_j^\dagger\}$ is the conditional probability that the message ρ_i will lead to the outcome $E_j=A_j^\dagger A_j$ and $p(j)=\Sigma_i p(j|i)$. Thus we now have

$$S(\rho_i\|\rho) \geq \sum_j p(j|i)[\ln p(j|i) - \ln p(j)].$$

Multiplying both sides by the (positive) p_i and summing over all i leads to the Holevo bound. ■

Since Holevo's result is one of the key results in quantum information theory, I present another simple way of understanding it via the quantum mutual information. This, of course, is only an additional motivation for the Holevo bound and by no means proves its validity. Namely, if Alice encodes the symbol (sym) i into the state (st) ρ_i, then the total state (sym+st) is

$$\rho_{sym+st}=\sum_i p_i|i\rangle\langle i|\otimes\rho_i,$$

where the kets $|i\rangle$ are orthogonal (we can think of these as representing different states of consciousness of Alice!). Bob now wants to learn about the symbols by distinguishing the states ρ_i. He cannot learn more about the symbols than is already stored in the correlations between the symbols and the message states. This as we know is given by the quantum mutual information

$$I(\rho_{sym+st}) = S(\mathrm{sym}) + S(\mathrm{st}) - S(\mathrm{sym+st}) = S\left(\sum_i p_i\rho_i\right) - \sum_i p_i S(\rho_i), \quad (42)$$

which is the same as the Holevo bound.

I would now like to derive the capacity of a classical communication channel from the Holevo bound. I follow the reasoning of Gorden (1964), who was, in fact, the first person to conjecture the Holevo bound. As I mentioned earlier, the Holevo bound itself contains the classical capacity of a classical channel as a special case. This, as we might expect, happens when all ρ_i's are diagonal in the same basis, i.e., they commute (classically all the states and observables commute because they can be simultaneously specified and measured, which is in contrast with quantum mechanics). Therefore density matrices are reduced to classical probability distributions. Let us call this basis the B representation, with orthonormal eigenvectors $|b\rangle$. Then the probability that the measurement of the symbol represented by ρ_i will yield the value b is just $\langle b|\rho_i|b\rangle$. I call this the conditional probability, $p_i(b)$, that if ρ_i was sent the result b was obtained. Now the Holevo bound is

$$\mathbf{C}=S(\rho)-\sum_i p_i S(\rho_i)=S(\rho)-S_B(\rho_i),$$

where $S_B(\rho_i)$ is the conditional entropy given by

$$S_B(\rho_i)=\sum_i p_i \sum_b \langle b|\rho_i|b\rangle \ln\langle b|\rho_i|b\rangle = \sum_i p_i S(\rho_i).$$

Thus the Holevo bound reduces itself to the Shannon mutual information between the commuting messages and the measurement in the B representation.

The usual rule of thumb for obtaining quantum information-theoretic quantities from their classical counterparts is by convention

$$\sum \rightarrow \mathrm{trace},$$

$$\sum p(a) \rightarrow \rho_A,$$

so that, for example, the Shannon entropy $S[p(a)] = -\Sigma_i p(a_i)\ln p(a_i)$ now becomes the von Neumann entropy $S(\rho_A)=-\mathrm{Tr}\rho_A\ln\rho_A$.

Example. As the first application of the Holevo bound I shall compute the channel capacity of a bosonic field, e.g., an electromagnetic field (for an excellent review, see Caves and Drummond, 1994). The message information will now be encoded into modes of frequency ω and average photon number $\overline{m}(\omega)$. The signal power is assumed to be S. The noise in the channel is quantified by the average number of excitations $\overline{n}(\omega)$ and is assumed to be independent of the signal (i.e., the power of signal and noise is additive). We saw that when there is no noise in the channel the Holevo bound is equal to the entropy of the average signal. In order to compute the capacity we need to maximize this entropy with the constraint that the total power (or energy) is fixed. It is well known that thermal states are those that maximize the entropy. We thus assume that both the noise and signal +noise are in thermal equilibrium and follow the usual Bose-Einstein statistics. The noise power is

$$N=\frac{\pi(kT)^2}{12\hbar}.$$

The power of the output of the channel (signal+noise) is

$$P=S+N=\frac{\pi(kT_e)^2}{12\hbar},$$

where T_e is the equilibrium temperature of signal +noise. Therefore it follows that

$$T_e=(12\hbar S/\pi k^2+T^2)^{1/2}.$$

The state of the noise in the mode ω is

$$\rho_N(\omega)=\sum_n \frac{1-e^{-\hbar\omega/kT}}{e^{\overline{n}(\omega)\hbar\omega/kT}}|n\rangle\langle n|,$$

while the state of the output is

$$\rho_{N+S}(\omega)=\sum_n \frac{1-e^{-\hbar\omega/kT_e}}{e^{\overline{n}(\omega)\hbar\omega/kT_e}}|n\rangle\langle n|.$$

The capacity of the channel is given by the Holevo bound, which is

$$C=\int_{-\infty}^{\infty}\{S[\rho_{S+N}(\omega)]-S[\rho_N(\omega)]\}d\omega$$

$$=\frac{\pi kT}{6\hbar\ln 2}\{[12\hbar S/\pi(kT)^2+1]^{1/2}-1\}. \quad (43)$$

The integration is there to take into account all the modes of the field. Let us look at the two extreme limits of this capacity. In the high-temperature limit we obtain the classical capacity

$$C_C=\frac{S}{kT\ln 2}, \quad (44)$$

a result derived by Shannon and Weaver (1949). This states that in order to communicate one bit of information with this setup we need exactly $kT\ln 2$ amount of energy. In the low-temperature limit, on the other hand, quantum effects become important and the capacity becomes independent of T:

$$C_Q=\frac{\sqrt{\pi}}{\ln 2}\left\{\frac{S}{\hbar}\right\}^{1/2}, \quad (45)$$

a result which was derived by Stern (1960), Lebedev and Levitin (1963), Gordon (1964), and Yamamoto and Haus (1986), among others. Note also the appearance of Planck's constant, which is a key feature of quantum mechanics. If we wish to communicate one bit of information in this limit we need only $\hbar/\pi(\ln 2)\sim 10^{-34}$ J of energy. This is significantly less than the corresponding energy in the classical limit. Let us now compare the classical and quantum capacity limits to the total energy of N harmonic oscillators (bosons) in the same two limits. In the high-temperature limit the equipartition theorem is applicable and the total energy is $3NkT$ (i.e., it depends on temperature). In the low-temperature limit all the harmonic oscillators settle down to the ground state so that the total energy becomes $N\hbar\omega/2$ (i.e., it is independent of temperature and we see the quantum dependence through Planck's constant $\hbar$).

B. Schumacher's compression

The optimal communication through a noiseless channel using pure states is equivalent to data compression. We saw in Eq. (3) that the limit to the classical data compression is given by the entropy of the data's probability distribution. We would thus guess that the limit to quantum data compression is given by the von Neumann entropy of the set of states being compressed. This, in fact, turns out to be a correct guess, as was first proven by Schumacher (1995). So, Alice now encodes letters of her classical message into pure quantum states and sends these to Bob. For example, if $a\rightarrow|\psi_a\rangle$ and $b\rightarrow|\psi_b\rangle$, then Alice's message aab will be sent to Bob as the sequence of pure quantum states $|\psi_a\rangle|\psi_a\rangle|\psi_b\rangle$.

The exact problem can be phrased in the following equivalent fashion: suppose a quantum source randomly prepares different qubit states $|\psi_i\rangle$ with the corresponding probabilities p_i. A random sequence of n such states is produced. By how much can this be compressed, i.e., how many qubits do we really need to encode the original sequence (in the limit of large n)? First of all, the total density matrix is

$$\rho=\sum_i p_i|\psi_i\rangle\langle\psi_i|.$$

This matrix can now be diagonalized as

$$\rho=\sum_i r_i|r_i\rangle\langle r_i|,$$

where r_i and $|r_i\rangle$ are the eigenvectors and eigenvalues. This decomposition is, of course, indistinguishable from the original one (or any other decomposition for that matter). Thus we can think about compression in this new basis, which is easier as it behaves completely classically (since $\langle r_i|r_j\rangle=\delta_{ij}$). We can therefore invoke results from the previous section on classical typical sequences and conclude that the limit to compression is $n(-\Sigma_i r_i\ln r_i)$, i.e., n qubits can be encoded into $nS(\rho)$ qubits. No matter how the states are generated, as long as the total state is described by the same density matrix ρ its compression limit is its von Neumann entropy. This protocol and result will be very important when we discuss entanglement measures in the following section.

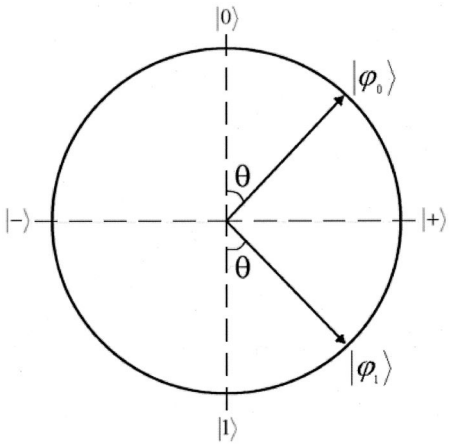

FIG. 4. Two nonorthogonal states on the Bloch sphere that are used to encode a message. The overlap between them is $\sin\theta$; the smaller the overlap, the more the total message can be compressed. In terms of information, the less distinguishable the states (i.e., the smaller the overlap), the less information they carry.

Example. Suppose that Alice encodes her bit into states $|\Psi_0\rangle=\cos(\theta/2)|0\rangle+\sin(\theta/2)|1\rangle$ and $|\Psi_1\rangle=\sin(\theta/2)|0\rangle+\cos(\theta/2)|1\rangle$ with $p_0=p_1=1/2$ (see Fig. 4). Classically it is not possible to compress a source that generates 0 and 1 with equal probability. Quantum mechanically, however, compression can be achieved not only by the nature of the probability distribution but also due to the nonorthogonality of the states encoding symbols of the message. In our example the overlap between the two states is $\langle\psi_0|\psi_1\rangle=\sin\theta$ and they are orthogonal only when $\theta=\pi$, in which case no compression is possible. Otherwise, the compression ratio is di-

rectly proportional to the overlap between the states. Suppose Alice's messages are only three qubits long. Then there are eight different possibilities, $|\Psi_0\Psi_0\Psi_0\rangle,\ldots,|\Psi_1\Psi_1\Psi_1\rangle$, which are all equally likely with 1/8 probability. In general these states will lie with a high probability within a subspace of the eight-dimensional Hilbert space. Let us call this likely subspace a "typical" subspace. Its orthogonal complement will be unlikely and hence called an "atypical" subspace. In order to find the typical and atypical subspaces we need to diagonalize the average signal,

$$\rho=\frac{1}{2}(|\Psi_0\rangle\langle\Psi_0|+|\Psi_1\rangle\langle\Psi_1|).$$

Its diagonal form is

$$\rho=\frac{1}{2}(1+\sin\theta)|+\rangle\langle+|+\frac{1}{2}(1-\sin\theta)|-\rangle\langle-|,$$

where $|\pm\rangle=|0\rangle\pm|1\rangle$. Now we look at the probabilities that each of the eight messages will lie along the new orthogonal basis $|+++\rangle,\ldots,|---\rangle$ of the Hilbert space of three qubits:

$$|\langle+++|\psi^{\otimes 3}\rangle|^2=[\cos(\theta/2)+\sin(\theta/2)]^6,$$

$$|\langle++-|\psi^{\otimes 3}\rangle|^2=[\cos(\theta/2)+\sin(\theta/2)]^4+[\cos(\theta/2)-\sin(\theta/2)]^2,$$

$$|\langle+--|\psi^{\otimes 3}\rangle|^2=[\cos(\theta/2)+\sin(\theta/2)]^2+[\cos(\theta/2)-\sin(\theta/2)]^4,$$

$$|\langle---|\psi^{\otimes 3}\rangle|^2=[\cos(\theta/2)-\sin(\theta/2)]^6,$$

where $|\psi^{\otimes 3}\rangle$ represents any three-qubit sequence of $|\psi_0\rangle$ and $|\psi_1\rangle$. In addition, all the probabilities for $|++-\rangle,|+-+\rangle,|-++\rangle$ are equal and so are the probabilities for $|+--\rangle,|--+\rangle,|-+-\rangle$. Thus the above equation contains 64 probabilities in total. Suppose now that $\cos(\theta/2)\sim\sin(\theta/2)$. Then we see that the states containing two or more + become much more likely. This means that the message states are much more likely to be in this particular subspace. Therefore the compression would be as follows. First the source generates three qubits in some state. Then we project this message onto the typical subspace. If we are successful, this will lie in that four-dimensional typical subspace for which we need only two qubits rather than three. Otherwise, our projection will fail and the message will end up in the atypical subspace, in which case Alice does not compress it. The probability of ending up in the atypical space asymptotically goes to zero (the law of large numbers). Therefore in this example the limit to our compression is given by $-[1/2(1+\sin\theta)]\ln[1/2(1+\sin\theta)]-[1/2(1-\sin\theta)]\ln[1/2(1-\sin\theta)]$, which is of course the von Neumann entropy of ρ. The number of dimensions of the total Hilbert space's typical subspace is likewise in general equal to $e^{nS(\rho)}$.

Interestingly, if instead of pure states a quantum source generates mixed states ρ_i with probabilities p_i, then the best compression limit is in general unknown. We can, of course, use the above protocol to compress the sequence to the von Neumann entropy of the average signal, $S(\Sigma_i p_i\rho_i)$. However, in some cases it is known that a better compression can be achieved. The lower bound to compression is the Holevo bound, $S(\Sigma_i p_i\rho_i)-\Sigma_i p_i S(\rho_i)$, but it is not known whether this bound can in general be attained (see Horodecki, 1998).

Next we look at a protocol for classical communication that involves entanglement. At first sight this protocol seems to violate the Holevo bound on classical communication, i.e., that it is possible to communicate only one bit per single qubit. However, a closer inspection will show that this is not the case.

C. Dense coding

Now let us consider the case of dense coding, which was introduced by Bennett and Wiesner (1992). In this protocol entanglement plays a crucial role, and this will give us a first indication of the fact that entanglement can be quantified like any other resource, such as energy, for example. Alice and Bob initially share an entangled pair of qubits in some state W_0, which may be mixed. Alice then performs local unitary operations on her qubit to put this shared pair of qubits into any of the states W_0, W_1, W_2, or W_3. In general, Alice may use a completely arbitrary set of unitary operations to generate these states:

$$W_i=\mathbf{U_i}\otimes\mathbf{I}W_0\mathbf{U_i}^{\dagger}\otimes\mathbf{I}, \tag{46}$$

and the number of generated states is completely arbitrary. In the above equation, $\mathbf{U_i}$ acts on Alice's qubit and $\mathbf{I}$ acts on Bob's qubit. By sending her encoded qubit to Bob, Alice is essentially communicating with Bob using the states W_0, W_1, W_2, and W_3 as separate letters. The number of bits she can communicate to Bob using this procedure is thus bounded by the Holevo bound. Moreover, if some block coding is done on a large enough collection of qubits in addition to the dense coding, then the number of bits of information communicated is equal to the Holevo function. We shall thus take

$$\mathbf{C}=S(\rho)-\sum_i p_i S(\rho_i), \tag{47}$$

assuming that any additional necessary block coding will automatically be performed to supplement the dense coding. This coding is essential in order to achieve the capacity given by the Holevo bound in the asymptotic limit. [The fact that the bound is achievable follows from a complicated argument and cannot really be derived using the arguments presented in this review. Hausladen *et al.* (1996) have proved this for pure states, and Schumacher and Westmoreland (1997) and independently Holevo (1998) have proved it for mixed states.] Exactly the same assumption has been used by Hausladen *et al.* (1996) to calculate the capacity for dense coding in the case of pure letter states. Equations (46) and (47) define the most general version of dense coding, and I shall refer to this as completely general dense coding.

A simpler example of dense coding is the case in which the letter states are generated from the initial shared state W_0 by

$$W_0 = \mathbf{I}\otimes\mathbf{I} W_0 \mathbf{I}\otimes\mathbf{I}, \tag{48}$$

$$W_1 = \sigma_\mathbf{1}\otimes\mathbf{I} W_0 \sigma_\mathbf{1}\otimes\mathbf{I}, \tag{49}$$

$$W_2 = \sigma_\mathbf{2}\otimes\mathbf{I} W_0 \sigma_\mathbf{2}\otimes\mathbf{I}, \tag{50}$$

$$W_3 = \sigma_\mathbf{3}\otimes\mathbf{I} W_0 \sigma_\mathbf{3}\otimes\mathbf{I}. \tag{51}$$

In the above set of equations, the first operator of the combination $\sigma_\mathbf{i}\otimes\mathbf{I}$ acts on Alice's qubit and the second operator acts on Bob's qubit. I shall refer to this case [i.e., in which the letter states are generated by Eqs. (48)–(51)] as simply general dense coding. The generality present in general dense coding is that Alice is allowed to prepare the different letter states with unequal probabilities.

In the more special case in which Alice not only generates the four letter states according to Eqs. (48)–(51) but also does so with equal probability, the ensemble is given by

$$W = \frac{1}{4}\sum_{i=0}^{3} W_i \tag{52}$$

and the capacity becomes

$$\mathbf{C} = \frac{1}{4}\sum_{i=0}^{3} S(W_i \| W). \tag{53}$$

I shall call this simplest case special dense coding. Among all the possible ways of doing general dense coding, special dense coding is the optimal way to communicate when W_0 is a pure state (Bose, Plenio, and Vedral, 2000) or a Bell diagonal state.

Now I derive the most general bound on completely general dense coding (Bowen, 2001). Furthermore, this bound can be attained by the same protocol as special dense coding (Bowen, 2001). The proof is achieved by first finding an upper bound to the capacity for completely general dense coding and then showing that special dense coding actually saturates this bound. Suppose that the initial state of Alice and Bob is ρ_{AB}. Then we have

$$\begin{aligned}\mathbf{C} &= \max S\left(\sum_k p_k (U^k\otimes I)\rho_{AB}[(U^k)^\dagger\otimes I]\right) \\ &\quad - \sum_k p_k S\{(U^k\otimes I)\rho_{AB}[(U^k)^\dagger\otimes I]\} \\ &= \max S(\rho'_{AB}) - S(\rho_{AB}) \leqslant S(\rho'_A) + S(\rho'_B) - S(\rho_{AB}) \\ &\leqslant 1 + S(\rho_B) - S(\rho_{AB}).\end{aligned} \tag{54}$$

Since this bound is achievable as shown by Bowen (2001), the capacity for completely general dense coding is given by Eq. (54).

I shall now restrict my attention to a calculation of **C** for pure letter states. Consider the initial shared pure state W_0 to be

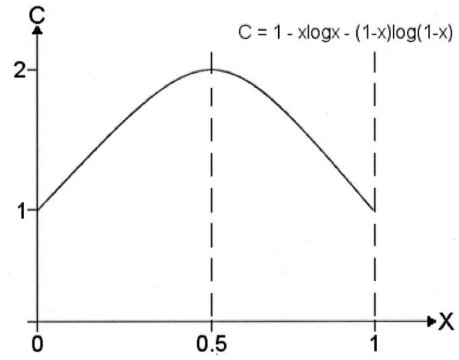

FIG. 5. The dependence of capacity for dense coding for pure states $a|00\rangle + b|11\rangle$ as a function of the Schmidt coefficient $x = |a|^2$. When the state is disentangled, i.e., when either $a=0$ or $b=0$, the capacity becomes 1 bit per qubit, the same as the classical capacity.

$$|\psi_0\rangle = (a|00\rangle + b|11\rangle). \tag{55}$$

Then, according to Eqs. (48)–(51), the other letter states are given by

$$|\psi_1\rangle = (a|10\rangle + b|01\rangle), \tag{56}$$

$$|\psi_2\rangle = -i(a|10\rangle - b|01\rangle), \tag{57}$$

$$|\psi_3\rangle = (a|00\rangle - b|11\rangle), \tag{58}$$

from which we obtain $W_i = |\psi_i\rangle\langle\psi_i|$. As all W_i are pure states we have

$$S(W_i) = 0. \tag{59}$$

Thus we have

$$\mathbf{C} = S(W). \tag{60}$$

I shall consider only the case of special dense coding as it is optimal. Thus the ensemble used is obtained from Eq. (52) to be

$$W = \frac{|a|^2}{2}|00\rangle\langle 00| + \frac{|b|^2}{2}|01\rangle\langle 01| + \frac{|a|^2}{2}|10\rangle\langle 10| + \frac{|b|^2}{2}|11\rangle\langle 11|.$$

Thus from Eq. (60) for the capacity **C** we get

$$\begin{aligned}\mathbf{C} &= -\left(|a|^2 \log\frac{|a|^2}{2} + |b|^2 \log\frac{|b|^2}{2}\right) \\ &= 1 - (|a|^2 \log|a|^2 + |b|^2 \log|b|^2).\end{aligned} \tag{61}$$

[Note that this agrees with Eq. (54), as for pure states the total entropy is zero.] This implies that a good measure of entanglement for a pure state of a system composed of two subsystems A and B can be given by the von Neumann entropy of the state of either of the subsystems. Let us call this measure the von Neumann entropy of entanglement and label it by E_v (Bennett, Bernstein, *et al.* 1996; Popescu and Rohrlich, 1997). Thus

$$E_v(|\psi\rangle\langle\psi|_{A+B}) = S[\mathrm{Tr}_A(|\psi\rangle\langle\psi|_{A+B})],$$

where Tr_A stands for partial trace over states of system A. Therefore, for all the states W_i,

$$E_v(W_i) = -(|a|^2 \log |a|^2 + |b|^2 \log |b|^2).$$

Thus

$$\mathbf{C} = 1 + E_v(W_i)$$

(see Fig. 5). We can prove that for pure states, special dense coding (using all alphabet states with equal *a priori* probability) is the optimal way to communicate among all possible ways of doing general dense coding [i.e., when the letter states are generated by Eqs. (48)–(51); Bose, Plenio, and Vedral, 2000]. It is important to understand that the amount of entanglement determines exactly how much information Alice can convey to Bob. Note that if there is no entanglement shared between them, then the amount of information is exactly one bit per Alice's qubit (which is what can be achieved classically after all). At the other extreme, when they share a maximally entangled state, the amount of information is two bits per Alice's qubit. This is an amount that no purely classical communication can achieve. However, while the von Neumann entropy is a good measure of entanglement for pure states (in fact, there are arguments that it is unique for pure states; Popescu and Rohrlich, 1997), it fails when we try to apply it to mixed states. A possibility is to follow the logic of the pure-state dense coding and call $S(\rho_B) - S(\rho_{AB})$ a measure of entanglement for mixed states as in Eq. (54). This measure has been called the "coherent information" and is used to describe information transmission through a noisy quantum channel (Barnum *et al.*, 1998). But is this measure consistent with other natural requirements for quantifying entanglement? This question will be addressed in the next section. Before this, we show that in order to delete a certain amount of correlation we need to increase the entropy of the environment by at least this amount. This is known as Landauer's erasure (Landauer, 1961; Toffoli, 1981; Bennett, 1988) and is seen to be linked directly to the relative entropy.

D. Relative entropy, thermodynamics, and information erasure

We have seen that communication essentially creates correlations between the sender and the receiver. Creating correlations is therefore very important in order to be able to convey any information. However, I should now like to talk about the opposite process—deleting correlations. Why would one want to do this? The reason is that one might want to correlate one system with another and might need to delete all its previous correlations to be able to store new ones. I should like to give a more physical statement of information erasure and link it to the notion of measurement. I shall therefore introduce two correlated parties—a system and an apparatus. The apparatus will interact with the system, thereby gaining a certain amount of information about it (the full quantum description of this process will be presented in Sec. V). Suppose that the apparatus needs to measure another system. We first need to delete information about the last system before we can make another measurement. The most general way of conducting erasure (resetting) of the apparatus is by employing a reservoir in thermal equilibrium at a certain temperature T. To erase the state of the apparatus we just throw it into the reservoir and introduce a new pure state. The entropy increase of the operation now consists of two parts. First, the state of the apparatus evolves to the state of the reservoir, and this entropy is added to the reservoir's entropy. Second, the rest of the reservoir changes its entropy due to this interaction, which is the difference in the apparatus's internal energy before and after the resetting (no work is done in this process). This quantum approach to equilibrium was also studied by Partovi (1989). A good model is obtained by imagining that the reservoir consists of a great number of systems (of the same "size" as the apparatus) all in the same quantum equilibrium state ω. Then the apparatus, which is in some state ρ, interacts with these reservoir systems one at a time. Each time there is an interaction, the state of the apparatus approaches more closely the state of the reservoir, while that single reservoir system also changes its state away from equilibrium. However, the systems in the bath are numerous, so that after a certain number of collisions the state of the apparatus will approach the state of the reservoir, while the reservoir will not change much since it is very large (this is equivalent to the Born-Markov approximation which leads to irreversible dynamics of the apparatus described here).

Bearing all this in mind, we now reset the apparatus by plunging it into a reservoir in thermal equilibrium (a Gibbs state) at temperature T. Let the state of the reservoir be

$$\omega = \frac{e^{-\beta H}}{Z} = \sum_j q_j |\varepsilon_j\rangle\langle\varepsilon_j|,$$

where $H = \Sigma_i \varepsilon_i |\varepsilon_i\rangle\langle\varepsilon_i|$ is the Hamiltonian of the reservoir, $Z = \mathrm{Tr}(e^{-\beta H})$ is the partition function, and $\beta^{-1} = kT$, where k is the Boltzmann constant. Now suppose that due to the measurement the entropy of the apparatus is $S(\rho)$ [and an amount $S(\rho)$ of information has been gained], where $\rho = \Sigma_i r_i |r_i\rangle\langle r_i|$ is the eigen expansion of the apparatus state. The total increase of entropy in the erasure (there are two parts as I argued above: change in the entropy of the apparatus and change in the entropy of the reservoir) is

$$\Delta S_{er} = \Delta S_{app} + \Delta S_{res}.$$

We immediately know that $\Delta S_{app} = S(\omega)$, since the state of the apparatus (no matter what state it was before) is now erased to become the same as that of the reservoir. However, the entropy change in the reservoir is the average over all states $|r_i\rangle$ of heat received by the reservoir divided by the temperature. This is minus the heat received by the apparatus divided by the temperature; the heat received by the apparatus is the internal energy after the resetting minus the initial internal energy $\langle r_i|H|r_i\rangle$. Thus

$$\Delta S_{res} = -\sum_k r_k \frac{\mathrm{Tr}(\omega H) - \langle r_k|H|r_k\rangle}{T}$$

$$= \sum_k \left(r_k \sum_j |\langle r_k|\varepsilon_j\rangle|^2 - q_k \right)(-\log q_k - \log Z)$$

$$= -\mathrm{Tr}(\rho-\omega)(\log\omega - \log Z) = \mathrm{Tr}(\omega-\rho)\log\omega.$$

Altogether we have an exact expression for the increase in entropy due to deletion:

$$\Delta S_{er} = -\mathrm{Tr}(\rho\log\omega).$$

This result (Vedral, 2000) generalizes Lubkin's result, which applies only when $[\rho,\omega]=0$. In general, however, the information gain is equal to $S(\rho)$, the entropy increase in the apparatus. This entropy increase is a maximum; the information between the system and apparatus is usually smaller, as in Eq. (42). Thus we see that

$$\Delta S_{er} = -\mathrm{Tr}(\rho\log\omega) \geqslant S(\rho) = I,$$

and Landauer's principle is confirmed [the inequality follows from the fact that the quantum relative entropy $S(\rho\|\omega) = -\mathrm{Tr}(\rho\log\omega) - S(\rho)$ is non-negative]. So the erasure is the least wasteful when $\omega=\rho$, in which case the entropy of erasure is equal to $S(\rho)$, the information gain. This is when the reservoir is in the same state as the state of the apparatus we are trying to erase. In this case we just have a state swap between the new pure state of the apparatus and the old state ρ which it replaces. Curiously enough, creating correlations is not costly in terms of the entropy of environment (such as when Alice and Bob communicate).

Landauer's erasure is a statement that is equivalent to the second law of thermodynamics. If we could delete information without increasing entropy, then we could construct a machine that completely converts heat into work with no other effect, which contradicts the second law. The opposite is also true. Namely, if we could convert heat into work with no other effect, then we could use this energy to delete information with no entropy increase (Landauer, 1961; Penrose, 1973; Toffoli, 1981; Bennett, 1988). Thus the relative entropy provides an interesting link between thermodynamics, information theory, and quantum mechanics [see also Brillouin's excellent book (Brillouin, 1956)].

I shall now show how Landauer's principle can be used to derive a limit to quantum data compression. The free energy lost in deleting information stored in a string of n qubits all is the state ρ is $n\beta^{-1}S(\rho)$. However, we could first compress this string and then delete the resulting information. The free-energy loss after compression is $m\beta^{-1}\log 2 = m\beta^{-1}$, where the string has been compressed to m qubits. The two free energies before and after compression should be equal if no information is lost during compression, i.e., if we wish to have maximal efficiency, and therefore $m/n = S(\rho)$ as shown previously (see Feynman, 1996). The equality is, of course, only achieved asymptotically.

So far we have seen that entropy plays a pivotal role in communication theory and data compression as a limit to both communication capacity and compression. It also quantifies the amount of entanglement in a pure bipartite state. Finally, it plays a thermodynamical role in characterizing the mixedness in a certain quantum state. This last role was first introduced by von Neumann. Now we go beyond the classical use of quantum states and address the question of how we can achieve quantum communication of quantum states.

IV. QUANTUM COMMUNICATION: QUANTUM USE

In this section the problem of entanglement quantification is analyzed. Previously we have seen that the reduced von Neumann entropy is a good measure of entanglement for two subsystems in a joint pure state (see also Bennett, Bernstein, *et al.*, 1996). This is a consequence of the Schmidt decomposition procedure introduced earlier and was exemplified by dense coding. However, for the mixed states of two subsystems, or for more than two subsystems, this procedure does not exist in general. Therefore it is not immediately clear how to understand and quantify correlations for these states. Initially we might think that Bell's inequalities (Clauser *et al.*, 1969; Bell, 1987; Redhead, 1987) would provide a good criterion for separating quantum correlations (entanglement) from classical correlations in a given quantum state. States that violate Bell's inequalities would be entangled and other states would be disentangled. However, while it is true that a violation of Bell's inequalities is a signature of quantum correlation, not all entangled states violate Bell's inequalities (Gisin, 1996). Therefore in order to completely separate quantum from classical correlations we need a different criterion.

I shall present here an approach that has proven to be very fruitful in understanding entanglement in general. It begins with a set of conditions that any reasonable measure of entanglement has to satisfy. I then discuss possible candidates based on these criteria.

A. Quantifying entanglement

In this section I shall mainly focus on understanding the entanglement of bipartite systems, i.e., systems containing only two subsystems. The term entanglement, or *versränkung* as it was originally called, was introduced by Schrödinger (1935) to emphasize the bizarre implications of quantum mechanics. The reason for studying bipartite entanglement is that it is the simplest and most basic kind of entanglement and is well understood at present. Starting from bipartite entanglement we can construct a theory that can be generalized to any number of systems.

To determine the basic properties that every "good" entanglement measure should satisfy (Vedral, Plenio, Rippin, *et al.*, 1997; Vedral and Plenio, 1998), we have to discuss what we actually mean when we say that something is "disentangled." By definition a bipartite state is disentangled if it can be written in the separable form $\rho_{AB} = \Sigma_i p_i \rho_i^A \otimes \rho_i^B$ (Werner, 1989). It is clear why we choose to define disentangled states in this manner:

these are the most general states that Alice and Bob can create by local operations and classical communication (LOCC). Thus these states contain no entanglement, as entanglement can be created only through global operations. All other states will be entangled to some degree. In addition, note that the set of all disentangled states is convex: a convex combination (mixture) of any two disentangled states is itself disentangled. This fact will be important when we quantify entanglement later.

The first question to answer is the following: When can a given matrix be written in a separable form? The necessary and sufficient condition is known in general in terms of positive (but not necessarily completely positive) maps (Horodecki *et al.*, 1996; Peres, 1996). Suppose that Λ is any positive map; then

$$I_A \otimes \Lambda_B \left(\sum_i p_i \rho_i^A \otimes \rho_i^B \right) = \sum_i p_i \rho_i^A \otimes \Lambda_B(\rho_i^B) \tag{62}$$

is always a positive operator. Remarkably, the converse is also true. If, for all positive maps Λ, the state $I_A \otimes \Lambda_B(\rho_{AB})$ is positive, then ρ_{AB} is separable (disentangled). Therefore, if we want to know whether a given state ρ_{AB} is entangled, we need to find a positive map whose action on B will result in a negative operator and hence not a physical state (Horodecki, 2001a). This condition is still not operational, since there is an infinite number of positive maps to search. In fact, there is no operational condition in general, but it exists only in some special cases. For example, for two qubits or a qubit and a qutrit (a three-level system), this condition simplifies to the following (Horodecki *et al.*, 1996; Peres, 1996): such a state is entangled if and only if a transposition of B results in a negative operator, i.e., $\rho_{AB}^{T_B} < 0$. The relationship between positive maps and entanglement is a very active field of research and I refer the interested reader to some papers investigating this issue: Bennett, DiVincenzo, Mor, *et al.* (1999); DiVincenzo *et al.* (2000); Kraus *et al.* (2000); Lewenstein *et al.* (2000). With this in mind, let us turn to quantifying entanglement.

The first property we need from an entanglement measure is that a disentangled state not have any quantum correlations. This gives rise to our first condition.

(E1) For any separable state σ the measure of entanglement should be zero, i.e.,

$$E(\sigma) = 0. \tag{63}$$

Note that we do not ask the converse, i.e., that if $E(\sigma) = 0$, then σ is separable. The reason for this will become clear below.

The next condition concerns the behavior of the entanglement under simple local unitary transformations. A local unitary transformation simply represents a change of the basis in which we consider the given entangled state. But a change of basis should not change the amount of entanglement that is accessible to us, because at any time we could just reverse the basis change (since unitary transformations are fully reversible).

(E2) For any state σ and any local unitary transformation, i.e., a unitary transformation of the form $U_A \otimes U_B$, the entanglement remains unchanged. Therefore

$$E(\sigma) = E(U_A \otimes U_B \sigma U_A^\dagger \otimes U_B^\dagger). \tag{64}$$

The third condition is the one that really restricts the class of possible entanglement measures. Unfortunately it is also the property that is usually the most difficult to prove for potential measures of correlation between two subsystems should increase under local operations on the subsystems separately. However, quantum entanglement is even more restrictive in that the total amount of entanglement cannot increase locally even with the aid of classical communication. Classical correlations, on the other hand, can be increased by the use of local operations and classical communication.

Example. Suppose that Alice and Bob share n uncorrelated pairs of qubits, for example, all in the state $|0\rangle$. Alice's computer then interacts with each of her qubits such that it randomly flips each qubit with probability 1/2. However, whenever a qubit is flipped, Alice's computer (classically) calls Bob's computer and informs it to do likewise. After this action on all the qubits, Alice and Bob end up sharing n (maximally) correlated qubits in the state $|00\rangle\langle 00| + |11\rangle\langle 11|$, i.e., whenever Alice's qubit is zero so is Bob's and whenever Alice's qubit is one so is Bob's. The state of each pair is mixed because Alice and Bob do not know whether their computers flipped their respective qubits or not.

We can always calculate the total amount of entanglement by summing up the entanglement of all systems after we have applied our local operations and classical communications.

(E3) Local operations, classical communication, and subselection cannot increase the expected entanglement, i.e., if we start with an ensemble in state σ and end up with probability p_i in subensembles in state σ_i then we shall have

$$E(\sigma) \geqslant \sum_i p_i E(\sigma_i), \tag{65}$$

where $\sigma_i = A_i \otimes B_i \sigma A_i^\dagger \otimes B_i^\dagger / p_i$ and $p_i = \mathrm{Tr}(A_i \otimes B_i \sigma A_i^\dagger \otimes B_i^\dagger)$. The form $A \otimes B$ shows that Alice and Bob perform their operation locally (i.e., Alice cannot affect Bob's system and vice versa). However, Alice's and Bob's operations can be correlated, as is manifested in the fact that they have the same index. It should be pointed out that although all the local operations and classi-

cal communication can be cast in the above product form, the opposite is not true: not all the operations of the product form can be executed locally (Bennett, Divincenzo, Fuchs, *et al.*, 1999). This means that the above condition is more restrictive than necessary, but this does not have any significant consequences as far as I am aware. An example of (E3) operation is the local addition of particles on Alice's and Bob's side. Note also that (E2) operations are a subset (special case) of (E3) operations.

The last condition is there to make sure that our measure is consistent with pure states.

(E4) Entanglement of a pure state is equal to the reduced von Neumann entropy.

The above conditions are natural and easy to understand physically. However, they can be reduced to simpler and more elementary conditions, which I now briefly discuss. Suppose that we ask that the measure of entanglement be

(1) weakly additive, i.e., $E(\rho\otimes\rho)=2E(\rho)$;
(2) continuous, i.e., if ρ is close to σ, then $E(\rho)$ is close to $E(\sigma)$.

Then, it can be shown (Popescu and Rohrlich, 1997; Vidal, 2000) that (E4) is a consequence of the weak additivity and continuity (providing we assume that the entanglement of a maximally entangled state is normalized to $\log 2$). In addition, in (E3) we use the most general local POVMs, but we know that these can be implemented by adding ancillas locally, performing a unitary transformation on the system and ancilla locally, and then tracing out the ancillas. So, (E2)–(E4) can be presented in a more elementary way, as was done by Vidal (2000). However, I chose to introduce entanglement measures via (E1)–(E4) as I think that they are more intuitive and capture the main ideas. Readers interested in further analysis of these conditions are advised to read Vidal (2000) and Horodecki *et al.* (2000).

Before I introduce different entanglement measures I should like to discuss the following question: What do we mean by saying that a state σ can be converted into another state ρ by local operations and classical communication? Strictly speaking, we mean that there exists an LOCC procedure that, given a sufficiently large number of copies n of σ, will convert them arbitrarily close to m copies of the state ρ; i.e.,

$$(\forall\epsilon>0)(\forall m\in N)(\exists n\in N;\exists\Phi\in\mathrm{LOCC}) \times\|\Phi(\sigma^{\otimes n})-\rho^{\otimes m}\|<\epsilon, \tag{66}$$

where $\|\sigma-\rho\|$ is some measure of distance (metric) on the set of density matrices. Now, if σ is more entangled than ρ, we expect that there is an LOCC procedure such that $m>n$; otherwise, we expect that we can have $n\leqslant m$. Measuring entanglement now reduces to finding an appropriate function on the set of states to order them according to their local convertibility. This is usually achieved by letting either σ or ρ be a maximally entangled state.

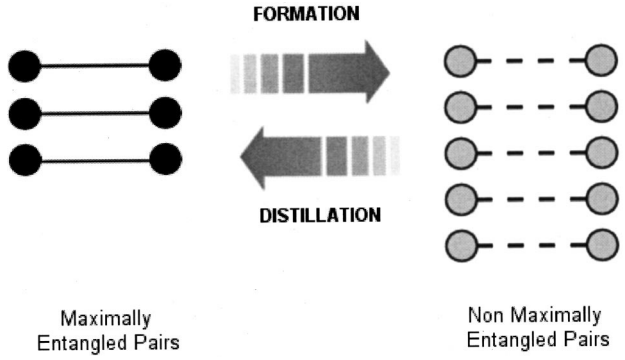

FIG. 6. The formation of entangled states: a certain number of maximally entangled pairs is manipulated by local operations and classical communication and converted into pairs in some state ρ. The asymptotic conversion ratio is known as the entanglement of formation. The converse of formation is distillation of entanglement. The asymptotic rate of converting pairs in state ρ into maximally entangled states is known as the entanglement of distillation. The two measures of entanglement are in general different, distillation being greater than or equal to formation. This surprising irreversibility of entanglement conversion is explained in the text as a consequence of the loss of classical information about the decomposition of ρ.

Entanglement of formation. I now introduce three different measures of entanglement, all of which obey (E1)–(E4). First I discuss the entanglement of formation of Bennett *et al.* (Bennett, Divincenzo, *et al.*, 1996). We define the entanglement of the formation of a state ρ by

$$E_F(\rho):=\min\sum_i p_i S(\rho_A^i), \tag{67}$$

where $S(\rho_A)=-\mathrm{Tr}\rho_A\ln\rho_A$ is the von Neumann entropy and the minimum is taken over all the possible realizations of the state, $\rho_{AB}=\Sigma_j p_j|\psi_j\rangle\langle\psi_j|$ with $\rho_A^i=\mathrm{Tr}_B(|\psi_i\rangle\langle\psi_i|)$. This measure satisfies (E1)–(E4). The basis of formation is that Alice and Bob would like to create an ensemble of n copies of the nonmaximally entangled state, ρ_{AB}, using only local operations, classical communication, and a number m of maximally entangled pairs (see Fig. 6). Entanglement of formation is the asymptotic conversion ratio, m/n, in the limit of infinitely many copies. The form of this measure given in Eq. (67) will be more transparent after the next subsection and the relationship between the entanglement of formation and other proposed measures will be analyzed in more detail below. It is worth mentioning that a closed form for this measure exists for two qubits (Wootters, 1998).

Related to this measure is the entanglement of distillation, also introduced by Bennett, Divincenzo, *et al.* (1996).

Entanglement of distillation. This measure defines the amount of entanglement of a state σ as the asymptotic proportion of singlets that can be distilled using a purification procedure (for a rigorous definition see Rains,

1999a, 1999b). This is the opposite process to that leading to the entanglement of formation (Fig. 6), although its value is generally smaller, implying that the formation of states is in some sense irreversible. The reason for this irreversibility will be explained in the next subsection. This measure fails to satisfy the converse of (E1), namely, for all disentangled states the entanglement of distillation is zero, but the converse is not true. There do exist states that are entangled, but from which no entanglement can be distilled. For this reason they are called *bound entangled states* (Horodecki *et al.*, 1998; see also DiVincenzo *et al.*, 2000). This is why the condition (E1) is not stated to be both necessary and sufficient.

Relative entropy of entanglement. I now introduce the final measure of entanglement, which was first proposed by Vedral, Plenio, Rippin, and Knight (1997). This measure is intimately related to the entanglement of distillation by providing an upper bound for it. If $\mathcal{D}$ is the set of all disentangled states, the measure of entanglement for a state σ is then defined as

$$E(\sigma) \equiv \min_{\rho \in \mathcal{D}} S(\sigma \| \rho), \tag{68}$$

where $S(\sigma\|\rho)$ is the quantum relative entropy. This measure, which I shall call the *relative entropy of entanglement*, tells us that the amount of entanglement in σ is its distance from the disentangled set of states. In statistical terms, as introduced in Sec. II, the more entangled a state is, the more it is distinguishable from a disentangled state (Vedral, Plenio, Jacobs, and Knight, 1997). To better understand all three measures of entanglement we need to introduce another quantum protocol that relies fundamentally on entanglement.

Another condition that might be considered intuitive for a measure of entanglement is *convexity*. That is, we might require that

$$E\left(\sum_i p_i \sigma^i\right) \leq \sum_i p_i E(\sigma^i).$$

This states that mixing cannot increase entanglement. For example, an equal mixture of two maximally entangled states $|00\rangle+|11\rangle$ and $|00\rangle-|11\rangle$ is a separable state and consequently contains no entanglement. I did not include convexity as a separate requirement for an entanglement measure as it is not completely independent from (E3). This is because (E3) and the strong additivity $[E(\rho\otimes\sigma)=E(\rho)+E(\sigma)]$ imply convexity,

$$n\sum_i p_i E(\rho_i) = E(\rho_1^{\otimes p_1 n}\rho_2^{\otimes p_2 n}\cdots\ \rho_N^{\otimes p_N n}) \geq E\left[\left(\sum_i p_i\rho_i\right)^{\otimes n}\right] = nE\left(\sum_i p_i\rho_i\right),$$

where the equalities follow from the strong additivity assumption and the inequality is a consequence of (E3). The symbol $\rho^{\otimes m}$ means that we have m copies of the state ρ. Nevertheless, it is interesting to point out that any convex measure that satisfies continuity and weak additivity has to be bounded from below by the entanglement of distillation and from above by the entanglement of formation (Horodecki *et al.*, 2000). We shall see that most entanglement measures can in fact be generated using the quantum relative entropy.

It is interesting to note that the relative entropy of entanglement does in fact satisfy both convexity and continuity (Donald and Horodecki, 1999) although not additivity (Vollbrecht and Werner, 2001). Furthermore, we can easily show that it is an upper bound to the entanglement of distillation. For any pure state $|\psi\rangle$, $\min_{\omega\in\mathcal{D}} S(\psi^{\otimes n}\|\omega) = \min_{\omega\in\mathcal{D}} -\langle\psi^{\otimes n}|\log\omega|\psi^{\otimes n}\rangle$. But the logarithmic function is concave, so that

$$\min_{\omega\in\mathcal{D}} -\langle\psi^{\otimes n}|\log\omega|\psi^{\otimes n}\rangle \geq \min_{\omega\in\mathcal{D}} -\log\langle\psi^{\otimes n}|\omega|\psi^{\otimes n}\rangle.$$

However, according to the recent result of Horodecki (Horodecki *et al.*, 1996), since ω is a disentangled state, its fidelity with the maximally entangled state cannot be larger than the inverse of the half dimension of that state, so that $\langle\psi^{\otimes n}|\omega|\psi^{\otimes n}\rangle \leq 1/2^n$. Thus

$$\min_{\omega\in\mathcal{D}} S(\psi^{\otimes n}\|\omega) \geq -\log(1/2^n) = n. \tag{69}$$

But we know that this minimum is achievable by the state $\omega=\rho^{\otimes n}$, where ρ is obtained from ψ by removing the off-diagonal elements in the Schmidt basis. Consequently, if we are starting with n copies of state σ and obtaining m copies of ψ by local operations and classical communication, then

$$D = \frac{m}{n} = \frac{1}{n}\min_{\omega\in\mathcal{D}} S(\psi^{\otimes m}\|\omega) \leq \frac{1}{n}\min_{\omega\in\mathcal{D}} S(\sigma^{\otimes n}\|\omega),$$

where the equality follows from Eq. (69) and the inequality from the fact that the relative entropy is nonincreasing under LOCC [strictly speaking, $D=\lim_{n\to\infty}(m/n)$ and, of course, m is a function of n, $m=m(n)$]. Thus the distillable entanglement is bounded from above by the relative entropy of entanglement.

A similar argument can be given to show that the relative entropy of entanglement is bounded from above by the entanglement of formation (Vedral and Plenio, 1998). Since most of the measures of entanglement can be derived from the relative entropy they will possess similar properties. In order to see this, we first need to introduce quantum teleportation.

B. Teleportation

Let us begin by describing quantum teleportation in the form originally proposed by Bennett *et al.* (1993). Suppose that Alice and Bob, who are distant from each other, wish to implement a teleportation procedure. Initially they need to share a maximally entangled pair of qubits. This means that if Alice and Bob each have one qubit, then the joint state may, for example, be

$$|\Psi_{AB}\rangle = (|0_A\rangle|0_B\rangle + |1_A\rangle|1_B\rangle)/\sqrt{2}, \tag{70}$$

where the first ket (with subscript A) belongs to Alice and second (with subscript B) to Bob. Note that this state is maximally entangled and is different from a sta-

tistical mixture $(|00\rangle\langle 00|+|11\rangle\langle 11|)/2$, which is the most correlated state allowed by classical physics.

Now suppose that Alice receives a qubit in an unknown state $|\Phi\rangle=a|0\rangle+b|1\rangle$ and she wants to teleport it to Bob. The state has to be unknown to her because otherwise she can just phone Bob up and tell him all the details of the state, and he can then recreate it on a particle that he possesses. Given that Alice does not know the state, she cannot measure it to obtain all the necessary information to specify it. If she could, this would lead to a violation of the uncertainty principle. Therefore she has to resort to using the state $|\Psi_{AB}\rangle$ that she shares with Bob to transfer her state to him without actually learning this state. This procedure is what we mean by quantum teleportation.

I first write out the total state of all three qubits,

$$|\Phi_{AB}\rangle := |\Phi\rangle|\Psi_{AB}\rangle = (a|0\rangle+b|1\rangle)(|00\rangle+|11\rangle)/\sqrt{2}.$$

However, the above state can be conveniently written in a different basis,

$$\begin{aligned}|\Phi_{AB}\rangle &= (a|000\rangle+a|011\rangle+b|100\rangle+b|111\rangle)/\sqrt{2}\\ &= \frac{1}{2}[|\Phi^+\rangle(a|0\rangle+b|1\rangle)+|\Phi^-\rangle(a|0\rangle-b|1\rangle)\\ &\quad +|\Psi^+\rangle(a|1\rangle+b|0\rangle)+|\Psi^-\rangle(a|1\rangle-b|0\rangle)],\end{aligned}$$

where

$$|\Phi^+\rangle=(|00\rangle+|11\rangle)/\sqrt{2}, \tag{71}$$

$$|\Phi^-\rangle=(|00\rangle-|11\rangle)/\sqrt{2}, \tag{72}$$

$$|\Psi^+\rangle=(|01\rangle+|10\rangle)/\sqrt{2}, \tag{73}$$

$$|\Psi^-\rangle=(|01\rangle-|10\rangle)/\sqrt{2}, \tag{74}$$

form an orthonormal basis of Alice's two qubits (remember that the first two qubits belong to Alice and the last qubit belongs to Bob). The above basis is frequently called the Bell basis. This is a very useful way of writing the state of Alice's two qubits and Bob's single qubit because it displays a high degree of correlation between Alice's and Bob's parts: for every state of Alice's two qubits (i.e., $|\Phi^+\rangle,|\Phi^-\rangle,|\Psi^+\rangle,|\Psi^-\rangle$) there is a corresponding state of Bob's qubit. In addition, the state of Bob's qubit in all four cases looks very much like the original qubit that Alice has to teleport to Bob. It is now straightforward to see how to proceed with the teleportation protocol (Bennett *et al.*, 1993).

(1) Upon receiving the unknown qubit in state $|\Phi\rangle$ Alice performs projective measurements on her two qubits in the Bell basis. This means that she will obtain one of the four Bell states randomly and with equal probability.

(2) Suppose Alice obtains the state $|\Psi^+\rangle$. Then the state of all three qubits (Alice+Bob) collapses to the following state:

$$|\Psi^+\rangle(a|1\rangle+b|0\rangle)$$

(the last qubit belongs to Bob as usual). Alice now has to communicate the result of her measurement to Bob (over the phone, for example). The point of this communication is to inform Bob how the state of his qubit now differs from the state of the qubit Alice was holding before the Bell measurement.

(3) Now Bob has to apply a unitary transformation on his qubit that simulates a logical NOT operation: $|0\rangle\rightarrow|1\rangle$ and $|1\rangle\rightarrow|0\rangle$. He thereby transforms the state of his qubit into the state $a|0\rangle+b|1\rangle$, which is precisely the state that Alice had to teleport to him initially. This completes the protocol. It is easy to see that if Alice obtained some other Bell state, then Bob would have to apply some other simple operation to complete the teleportation. They can be represented by the Pauli spin matrices.

An important fact to observe in the above protocol is that all the operations (Alice's measurements and Bob's unitary transformations) are local in nature. This means that there is never any need to perform a (global) transformation or measurement on all three qubits simultaneously, which is what allows us to call the above protocol a genuine teleportation. It is also important that the operations that Bob performs are independent of the state that Alice tries to teleport to him. Note also that the classical communication from Alice to Bob in step (2) above is crucial because otherwise the protocol would be impossible to execute. (There is a deeper reason for this: if we could perform teleportation without classical communication, then Alice could send messages to Bob faster than the speed of light; see, for example, Vedral, Rippin, and Plenio, 1997.)

It is important to observe that the initial state to be teleported is destroyed immediately after Alice's measurement, i.e., it becomes maximally mixed of the form $(|0\rangle\langle 0|+|1\rangle\langle 1|)/2$. This has to happen since otherwise Alice and Bob would end up with two qubits in the same state, effectively cloning an unknown quantum state, which is impossible by the laws of quantum mechanics. This is the no-cloning theorem of Wootters and Zurek (1982), which is a simple consequence of the linearity of quantum dynamical laws. We also see that at the end of the protocol the quantum entanglement of $|\Psi_{AB}\rangle$ is completely destroyed. Does this have to be the case in general or might we save that state at the end (perhaps by performing a different teleportation protocol)? The answer is yes, the state must be destroyed (Plenio and Vedral, 1998), because if this were not the case, then entanglement could increase under local operations and classical communication, which as we have seen is prohibited by definition.

Teleportation has been experimentally performed in three different setups (Bouwmeester *et al.*, 1997; Boschi *et al.*, 1998; Furusawa *et al.*, 1998). It will now be used to link the three measures of entanglement. I shall show that all the different measures of entanglement can be understood as special cases of the relative entropy of entanglement (Henderson and Vedral, 2000). This unification relies on adding an ancilla, which I shall call a memory system and which will help us keep track of the various decompositions of a given bipartite density ma-

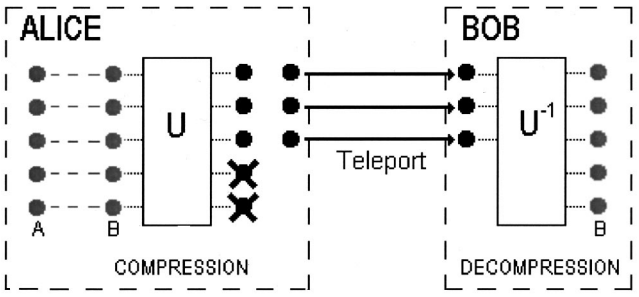

FIG. 7. Formation of a state by local operations and classical communication and with the help of teleportation. First, Alice creates the joint state of subsystems A and B locally. Then, she performs quantum data compression on the subsystem B and teleports the compressed state to Bob. Finally, Bob decompresses the received state. Hence Alice and Bob end up sharing the joint state of A and B initially prepared by Alice.

trix. How much access is available to this memory determines which measure of entanglement is used.

C. Measures of entanglement from relative entropy

Suppose that Alice and Bob share a state described by the density matrix ρ_{AB}. The state ρ_{AB} has an infinite number of different decompositions $\varepsilon = \{|\psi^i_{AB}\rangle\langle\psi^i_{AB}|, p_i\}$ into pure states $|\psi^i_{AB}\rangle$, with probabilities p_i. We denote the mixed state ρ_{AB} written in decomposition ε by

$$\rho^{\varepsilon}_{AB} = \sum_i p_i |\psi^i_{AB}\rangle\langle\psi^i_{AB}|. \tag{75}$$

As we have seen, measures of entanglement are associated with the formation and distillation of pure and mixed entangled states. The known relationships between the different measures of entanglement for mixed states are $E_D(\rho_{AB}) \leqslant E_{RE}(\rho_{AB}) \leqslant E_F(\rho_{AB})$ (Vedral and Plenio, 1998). Equality holds for pure states, in which all the measures reduce to the von Neumann entropy, $S(\rho_A) = S(\rho_B)$.

Formation of an ensemble of n nonmaximally entangled *pure* states, $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$, is achieved by the following protocol. Alice first prepares the states she would like to share with Bob locally. She then uses the Schumacher compression (Jozsa and Schumacher, 1994; Schumacher, 1995), to compress subsystem B into $nS(\rho_B)$ states. Subsystem B is then teleported to Bob using $nS(\rho_B)$ maximally entangled pairs. Bob decompresses the states he receives and so ends up sharing n copies of ρ_{AB} with Alice. The entanglement of formation is therefore $E_F(\rho_{AB}) = S(\rho_B)$. For pure states, this process requires no classical communication in the asymptotic limit (Lo and Popescu, 1999). The reverse process of distillation is accomplished using the Schmidt projection method (Bennett, Bernstein, *et al.*, 1996), which allows $nS(\rho_B)$ maximally entangled pairs to be distilled in the limit as n becomes very large. No classical communication between the separated parties is required. Therefore pure states are fully interconvertible in the asymptotic limit.

The situation for mixed states is more complex. When any mixed state, denoted by Eq. (75), is created, it may be imagined to be part of an extended system whose state is pure. The pure states $|\psi^i_{AB}\rangle$ in the mixture may be regarded as correlated to orthogonal states $|m_i\rangle$ of a memory M. The extended system is in the pure state $|\psi_{MAB}\rangle = \Sigma_i \sqrt{p_i}|m_i\rangle|\psi^i_{AB}\rangle$. If we have no access to the memory system, we trace over it to obtain the mixed state in Eq. (75). In fact, the lack of access to the memory is of a completely general nature. It may be due to interaction with another inaccessible system, or it may be due to an intrinsic loss of information. The results I shall present are universally valid and do not depend on the nature of the information loss. We shall see that the amount of entanglement involved in the different entanglement manipulations of mixed states depends on the accessibility of the information in the memory at different stages. Note that a unitary operation on $|\psi_{MAB}\rangle$ will convert it into another pure state $|\phi_{MAB}\rangle$ with the same entanglement, and tracing over the memory yields a different decomposition of the mixed state. Reduction of the pure state to the mixed state may be regarded as due to a projection-valued measurement on the memory with operators $\{E_i = |m_i\rangle\langle m_i|\}$.

Consider first the protocol of formation by means of which Alice and Bob come to share an ensemble of n mixed states ρ_{AB} as in Fig. 7. Alice first creates the mixed states locally by preparing a collection of n states in a particular decomposition, $\varepsilon = \{|\psi^i_{AB}\rangle\langle\psi^i_{AB}|, p_i\}$, by making np_i copies of each pure state $|\psi^i_{AB}\rangle$. At the same time we may imagine a memory system entangled with the pure states to be generated, which keeps track of the identity of each member of the ensemble. I consider first the case in which the states of subsystems A and B together with the memory are pure. Later I shall consider the situation in which Alice's memory is decohered. There are then three ways for her to share these states with Bob. First of all, she may simply compress subsystem B to $nS(\rho_B)$ states and teleport these to Bob using $nS(\rho_B)$ maximally entangled pairs. The choice of which subsystem to teleport is made so as to minimize the amount of entanglement required, so that $S(\rho_B) \leqslant S(\rho_A)$. The teleportation in this case would require no classical communication in the asymptotic limit, just as for pure states (Lo and Popescu, 1999). The whole system that is created by this process is an ensemble of pure states $|\psi_{MAB}\rangle$, where subsystems M and A are on Alice's side and subsystem B is on Bob's side. In terms of entanglement resources, however, this process is not the most efficient way for Alice to send the states to Bob. She may do it more efficiently by using the memory system of $|\psi_{MAB}\rangle$ to identify blocks of np_i members in each pure state $|\psi^i_{AB}\rangle$ and applying compression to each block to give $np_iS(\rho^i_B)$ states. Then the total number of maximally entangled pairs required to teleport these

states to Bob is $n\Sigma_i p_i S(\rho_B^i)$, which is clearly less than $nS(\rho_B)$, by concavity of the entropy.

The amount of entanglement required clearly depends on the decomposition of the mixed state ρ_{AB}. In order to decompress these states, Bob must also be able to identify which members of the ensemble are in which state. Therefore Alice must also send him the memory system. She now has two options. She may either teleport the memory to Bob, which would use more entanglement resources, or she may communicate the information in the memory classically, with no further use of entanglement. When Alice uses the minimum entanglement decomposition, $\varepsilon=\{|\psi_{AB}^i\rangle\langle\psi_{AB}^i|,p_i\}$, this process, originally introduced by Bennett, DiVincenzo, *et al.* (1996), makes the most efficient use of entanglement, consuming only the entanglement of formation of the mixed state, $E_F(\rho_{AB})=\Sigma_i p_i S(\rho_B^i)$.

We may think of the classical communication between Alice and Bob in one of two equivalent ways. Alice may either measure the memory locally to decohere it and then send the result to Bob classically, or she may send the memory through a completely decohering quantum channel. Since Alice and Bob have no access to the channel, the state of the whole system created by this process is the mixed state

$$\rho_{ABM}^{\varepsilon}=\sum_i p_i|\psi_{AB}^i\rangle\langle\psi_{AB}^i|\otimes|m_i\rangle\langle m_i|, \tag{76}$$

where Bob is classically correlated to the AB subsystem. Bob is then able to decompress his states using the memory to identify members of the ensemble.

Once the collection of n pairs is shared between Alice and Bob, it is converted into an ensemble of n mixed states ρ_{AB} by destroying access to the memory which contains the information about the state of any particular member of the ensemble. It is the loss of this information that is responsible for the fact that entanglement of distillation is lower than entanglement of formation, since it is not available to parties carrying out the distillation. If Alice and Bob, who do have access to the memory, were to carry out the distillation, they could obtain as much entanglement from the ensemble as was required to form it. In the case in which Alice and Bob share an ensemble of the pure state $|\psi_{MAB}\rangle$, they would simply apply the Schmidt projection method (Bennett, Bernstein, *et al.*, 1996). The relative entropy of entanglement gives the upper bound to distillable entanglement, $E_{RE}(|\psi_{(MA):B}\rangle\langle\psi_{(MA):B}|)=S(\rho_B)$, which is the same as the amount of entanglement required to create the ensemble of pure states, as described above. Here M, A, and B are spatially separated subsystems on which joint operations may not be performed. In my notation, I use a colon to separate the local subsystems.

On the other hand, if Alice uses the least entanglement for producing an ensemble of the mixed state ρ_{AB}, together with classical communication, the state of the whole system is an ensemble of the mixed state ρ_{ABM}^{ε}, and the process is still reversible. Because of the classical correlation to the states $|\psi_{AB}^i\rangle$, Alice and Bob may identify blocks of members in each pure state $|\psi_{AB}^i\rangle$ and apply the Schmidt projection method to them, giving $np_iS(\rho_B^i)$ maximally entangled pairs, and hence a total entanglement of distillation of $\Sigma_i p_i S(\rho_B^i)$. The relative entropy of entanglement again quantifies the amount of distillable entanglement from the state ρ_{ABM}^{ε} and is given by $E_{RE}(\rho_{A:(BM)}^{\varepsilon})=\min_{\sigma_{ABM}\in D}S(\rho_{ABM}^{\varepsilon}\|\sigma_{ABM})$. The disentangled state that minimizes the relative entropy is $\sigma_{ABM}=\Sigma_i p_i\sigma_{AB}^i\otimes|m_i\rangle\langle m_i|$, where σ_{AB}^i is obtained from $|\psi_{AB}^i\rangle\langle\psi_{AB}^i|$ by deleting the off-diagonal elements in the Schmidt basis. This is the minimum because the state ρ_{MAB} is a mixture of the orthogonal states $|m_i\rangle|\psi_{AB}^i\rangle$, and for a pure state $|\psi_{AB}^i\rangle$, the disentangled state that minimizes the relative entropy is σ_{AB}^i. The minimum relative entropy of the extended system is then

$$S(\rho_{ABM}^{\varepsilon}\|\sigma_{ABM})=\sum_i p_i S(\rho_B^i).$$

This relative entropy, $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$, has previously been called the *entanglement of projection* (Garisto and Hardy, 1999) because the measurement on the memory projects the pure state of the full system into a particular decomposition. The minimum of $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$ over all decompositions is equal to the entanglement of formation of ρ_{AB}. However, Alice and Bob may choose to create the state ρ_{AB} by using a decomposition with higher entanglement than the entanglement of formation. The maximum of $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$ over all possible decompositions is called the *entanglement of assistance* of ρ_{AB} (DiVincenzo *et al.*, 1999). Because $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$ is a relative entropy, it is invariant under local operations and nonincreasing under general operations, properties that are conditions for a good measure of entanglement (Vedral and Plenio, 1998). However, unlike $E_{RE}(\rho_{AB})$ and $E_F(\rho_{AB})$, it is not zero for completely disentangled states. In this sense, the relative entropy of entanglement, $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$, defines a class of entanglement measures interpolating between the entanglement of formation and the entanglement of assistance. Note that an upper bound for the entanglement of assistance, E_A, can be shown using concavity (DiVincenzo *et al.*, 1998) to be $E_A(\rho_{AB})\leqslant\min[S(\rho_A),S(\rho_B)]$. This bound can also be shown from the fact that the distillable entanglement from any decomposition, $E_{RE}(\rho_{A:(BM)}^{\varepsilon})\leqslant E_A(\rho_{AB})$, cannot be greater than the entanglement of the original pure state.

Note that here we are really creating a state $\rho^{\otimes n}=\rho\otimes\rho\cdots\rho$. The entanglement of formation of such a state is, strictly speaking, given by $E_F(\rho^{\otimes n})$, so the entanglement of formation per one single pair is $E_F(\rho^{\otimes n})/n$. It is at present not clear if this is the same as $E_F(\rho)$ in general, i.e., whether the entanglement of formation is additive. Bearing this in mind we continue our discussion, whose conclusions will not depend on the validity of the additivity assumption of the entanglement of formation (for more on this issue see, for example, Hayden *et al.*, 2001).

We may also derive relative entropy measures that interpolate between the relative entropy of entanglement and the entanglement of formation (Horodecki *et al.*, 2000) by considering nonorthogonal measurements on the memory. First of all, the fact that the entanglement of formation is in general greater than the upper bound for entanglement of distillation emerges as a property of the relative entropy, namely, it cannot increase under the local operation of tracing one subsystem [this is property (F2) of the quantum relative entropy given in Sec. II; Lindblad, 1974]:

$$E_F(\rho_{AB}) = \min_{\sigma_{ABM}\in\mathcal{D}} S(\rho_{ABM}\|\sigma_{ABM}) \geqslant \min_{\sigma_{AB}\in\mathcal{D}} S(\rho_{AB}\|\sigma_{AB}). \quad (77)$$

In general, the loss of the information in the memory may be regarded as a result of an imperfect classical channel. This is equivalent to Alice's making a nonorthogonal measurement on the memory and sending the result to Bob. In the most general case, $\{E_i = A_i A_i^+\}$ is a POVM [loosely speaking, this is a CP map as in Eq. (13), where all the individual outcomes are recorded] performed on the memory. The decomposition corresponding to this measurement is composed of mixed states, $\xi = \{q_i, \mathrm{Tr}_M(A_i \rho_{MAB} A_i^+)\}$, where $q_i = \mathrm{Tr}(A_i \rho_{MAB} A_i^+)$. The relative entropy of entanglement of the state ρ_{MAB}^{ξ}, when ξ is a decomposition of ρ_{AB} resulting from a nonorthogonal measurement on M, defines a class of entanglement measures interpolating between the relative entropy of entanglement and the entanglement of formation of the state ρ_{AB}. In the extreme case where the measurement gives no information about the state ρ_{AB}, $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$ becomes the relative entropy of entanglement of the state ρ_{AB} itself. In between, the measurement gives partial information. So far, I have shown that the measures interpolating between entanglement of assistance and entanglement of formation result from making orthogonal measurements on preparations of the pure state $|\psi_{MAB}\rangle$ in different bases. I note that they may equally be achieved by using the preparation associated with entanglement of assistance and making increasingly nonorthogonal measurements.

D. Classical information and quantum correlations

The loss of entanglement may be related to the loss of information in the memory. There are two stages at which distillable entanglement is lost. The first is in the conversion of the pure state $|\psi_{MAB}\rangle$ into a mixed state ρ_{ABM}. This happens because Alice uses a classical channel to communicate the memory to Bob. The second is due to loss of the memory M, taking the state ρ_{ABM} to ρ_{AB}. The amount of information lost may be quantified as the difference in mutual information between the respective states. Mutual information is a measure of correlations between the memory M and the system AB, giving the amount of information about AB that may be obtained from a measurement on M. The quantum mutual information between M and AB is defined as $I_Q(\rho_{M:(AB)}) = S(\rho_M) + S(\rho_{AB}) - S(\rho_{MAB})$. The mutual information loss in going from the pure state $|\psi_{MAB}\rangle$ to the mixed state in Eq. (76) is $\Delta I_Q = S(\rho_{AB})$. There is a corresponding reduction in the relative entropy of entanglement, from the entanglement of the original pure state, $E_{RE}(|\psi_{(MA):B}\rangle\langle\psi_{(MA):B}|)$, to the entanglement of the mixed state $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$ for all decompositions ε arising as the result of an orthogonal measurement on the memory. It is possible to prove, using the nonincrease of relative entropy under local operations, that when the mutual information loss is added to the relative entropy of entanglement of the mixed state $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$, the result is greater than the relative entropy of entanglement of the original pure state, $E_{RE}(|\psi_{(MA):B}\rangle\langle\psi_{(MA):B}|)$ (Henderson and Vedral, 2000). The strongest case, which occurs when $E_{RE}(\rho_{A:(BM)}^{\varepsilon}) = E_F(\rho_{AB})$, is

$$E_{RE}(|\psi_{(MA):B}\rangle\langle\psi_{(MA):B}|) \leqslant E_F(\rho_{AB}) + S(\rho_{AB}). \quad (78)$$

A similar result may be proved for the second loss, due to loss of the memory (Henderson and Vedral, 2000). Again the mutual information loss is $\Delta I_Q = S(\rho_{AB})$. The relative entropy of entanglement is reduced from $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$, for any decomposition ε resulting from an orthogonal measurement on the memory, to $E_{RE}(\rho_{AB})$, the relative entropy of entanglement of the state ρ_{AB} with no memory. When the mutual information loss is added to $E_{RE}(\rho_{AB})$, the result is greater than $E_{RE}(\rho_{A:(BM)}^{\varepsilon})$. In this case, the result is strongest for $E_{RE}(\rho_{A:(BM)}^{\varepsilon}) = E_A(\rho_{AB})$:

$$E_A(\rho_{AB}) \leqslant E_{RE}(\rho_{AB}) + S(\rho_{AB}). \quad (79)$$

Notice that if ρ_{AB} is a pure state, then $S(\rho_{AB}) = 0$, and equality holds. Inequalities (78) and (79) provide lower bounds for $E_F(\rho_{AB})$ and $E_{RE}(\rho_{AB})$, respectively. They are of a form typical of irreversible processes in that restoring the information in M is not sufficient to restore the original correlation between M and AB. In particular, they express that the loss of entanglement between Alice and Bob at each stage must be accompanied by an even greater reduction in mutual information between the memory and subsystems AB. The general result can be derived from Donald's equality (Donald, 1986, 1987). We have in general that for any σ and $\rho = \Sigma_i p_i \rho_i$ the following is true:

$$S(\rho\|\sigma) + \sum_i p_i S(\rho_i\|\rho) = \sum_i p_i S(\rho_i\|\sigma).$$

Suppose that $E(\rho) = S(\rho\|\sigma)$. Then, since $E(\rho_i) \leqslant S(\rho_i\|\sigma)$, we have the inequality

$$E(\rho) + \sum_i p_i S(\rho_i\|\rho) \geqslant \sum_i p_i E(\rho_i).$$

Thus the loss of entanglement in $\{p_i, \rho_i\} \to \rho$ is bounded from above by the Holevo information

$$\sum_i p_i E(\rho_i) - E(\rho) \leqslant \sum_i p_i S(\rho_i\|\rho). \quad (80)$$

This is a physically pleasing property of entanglement. It says that the amount of information lost always exceeds the lost entanglement, which indicates that entanglement stores only a part of the information; the rest, of course, is stored in classical correlations (see also Eisert *et al.*, 2000, who consider a similar problem, although not in the full generality of the above analysis).

In summary, the relative entropy of entanglement of the state ρ_{AB} depends only on the density matrix ρ_{AB} and gives an upper bound to the entanglement of distillation. The other measures of entanglement, which are given by relative entropies of an extended system, all depend on how the information in the memory is used or how the density matrix is decomposed. There are numerous decompositions of any bipartite mixed state into a set of states ρ_i with probability p_i. The average entanglement of states in each decomposition is given by the relative entropy of entanglement of the system extended by a memory whose orthogonal states are classically correlated to the states of the decomposition. This correlation records which state ρ_i any member of an ensemble of mixed states $\rho_{AB}^{\otimes n}$ is in. It is available to parties involved in the formation of the mixed state, but is not accessible to parties carrying out distillation. When the classical information is fully available, different decompositions give rise to different amounts of distillable entanglement, the highest being entanglement of assistance and the lowest, entanglement of formation. If access to the classical record is reduced, the amount of distillable entanglement is reduced. In the limit where no information is available, the upper bound to the distillable entanglement is given by the relative entropy of entanglement of the state ρ_{AB} itself, without the extension of the classical memory.

I close this section by discussing generalizations to more than two subsystems. First of all, it is not at all clear how to perform this in the case of entanglement of formation and distillation. The former just does not have a natural generalization and, for the latter, it is not clear what states we should be distilling when we have three or more parties. The relative entropy of entanglement on the other hand, does not suffer from this problem (Vedral, Plenio, Jacobs, and Knight, 1997; Vedral and Plenio, 1998). Its definition for N parties would be $E_{RE}(\sigma) := \min_{\rho \in D} S(\sigma \| \rho)$ where $\rho = \Sigma_i p_i \rho_1^i \otimes \rho_2^i \otimes \cdots \otimes \rho_N^i$.

I shall now use the knowledge we have gained of classical and quantum correlations to describe quantum computation. It will be seen, perhaps somewhat surprisingly, that classical correlations will play a more prominent role than quantum correlations in the speedup of certain quantum algorithms.

V. QUANTUM COMPUTATION

A quantum computer is a physical system that can accept input states which represent a coherent superposition of many different possible basis states and subsequently evolve them into a corresponding superposition of outputs. Computation, i.e., a sequence of unitary transformations, affects simultaneously each element of the superposition, generating a massive parallel data processing albeit within one piece of quantum hardware. In this way quantum computers can efficiently solve some problems that are believed to be intractable on classical computers (Deutsch and Josza, 1992; Bernstein and Vazirani, 1993; Simon, 1994). The best example is Shor's factorization algorithm (Shor, 1996); for overview of this algorithm see Ekert and Josza (1996). Therefore the advantage of a quantum computer lies in the exploitation of the phenomenon of superposition. The great importance of the quantum theory of computation is in the fact that it reveals the fundamental connections between the laws of physics and the nature of computation (Deutsch, 1998).

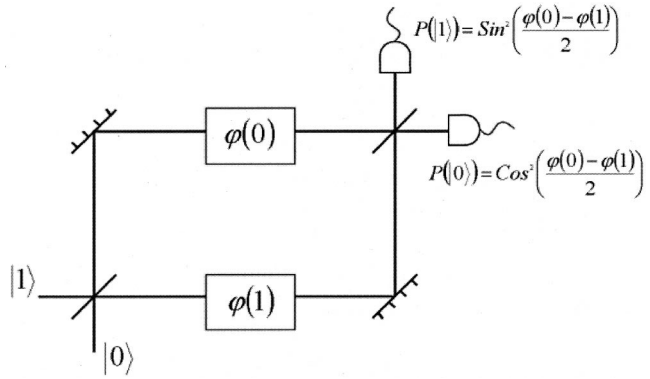

FIG. 8. The Mach-Zender interferometer. A photon is split at a beamsplitter and can take two different paths. In each of the paths we have a different phase introduced to the photon state, so that, after it encounters the second beamsplitter, the probabilities of detection in two branches have a sinusoidal dependence on the phase difference. In terms of quantum computation, the beamsplitter implements the Hadamard transform and the whole interferometer can be seen as implementing Deutsch's algorithm (see text for explanation).

In order to understand the efficiency of computer algorithms, we have to discuss the theory of computational complexity. I shall only mention the basics; a more detailed account can be found in the article of Papadimitriou (1995). Computational complexity concerns the difficulty of solving certain problems, such as the multiplication of two numbers, finding the minimum of a given function, and so on. Complexity theory divides problems into two basic categories:

(1) *Easy problems*: the time of computation T is a polynomial function of the size of the input l, for example, $T = c_n l^n + \cdots + c_1 l + c_0$, where the coefficients c are determined by the problem.

(2) *Hard problems*: the time of computation is an exponential function of the size of the input (for example, $T = 2^{cl}$, where c is problem dependent).

The size of the input is always measured in bits (qubits). For example, if we are to store the number 15, then we need 4 bits. In general, to store a number N we need about $l = \log N$, where the base of the logarithm is 2.

The division of problems into "easy" and "hard" is, of course, very rough. First of all, in computation, apart

from time, there are other resources that might matter, such as space, energy, and so on. If time grows polynomially but we require an exponentially increasing energy, then the problem is clearly difficult. In addition, suppose that the time complexity of one problem is $10^{10}n$ and that of another is $10^{-10}2^n$. Then for small n (say, $n=10$), the second algorithm, in spite of being exponential, is clearly more efficient. These two issues exemplify that the division into hard and easy problems is not without its own problems. However, this classification system is very simple to put into practice and does illuminate many different aspects of computational problems, which is why it is so widely used. I refer the reader to the book by Garey and Johnson (1979), which presents an introduction to hard problems and their detailed classification.

There is a great simplification in understanding quantum computation: a quantum computer is formally equivalent to a multiparticle Mach-Zender-like interferometer (Cleve *et al.*, 1997). I first present the simplest kind of interferometer in terms of its function as a simple computer. We see from Fig. 8 that the path of the photon is in fact a quantum bit in the sense that the photon can be in a superposition of the two paths. The first beam splitter acts as the unitary evolution $|0\rangle \rightarrow |0\rangle + |1\rangle$, which is known as the Hadamard gate. Next is the phase shift, which has the following effect:

$$|0\rangle \rightarrow e^{i\phi(0)}|0\rangle,$$

$$|1\rangle \rightarrow e^{i\phi(1)}|1\rangle.$$

At the end we have another beamsplitter and two detectors measuring contributions to the state $|0\rangle$ and $|1\rangle$. The corresponding probabilities of detection are

$$P_0 = \cos^2 \frac{\phi(0)-\phi(1)}{2},$$

$$P_1 = \sin^2 \frac{\phi(0)-\phi(1)}{2}.$$

If, for example, $\phi(0)=\phi(1)$, then only detector 0 will be registering counts. If, however, $\phi(0)=\phi(1)\pm\pi$, then only detector 1 will be registering counts. These two situations are basically identical to what is known as Deutsch's algorithm (Deutsch, 1985), the first algorithm to give an indication that quantum computers are more powerful than their classical counterparts. This algorithm has also been implemented experimentally in nuclear magnetic resonance (NMR) (Jones and Mosca, 1998).

A. Deutsch's algorithm

Deutsch's algorithm (Deutsch, 1985; see also Deutsch, 1998) is the simplest possible example that illustrates the advantages of quantum computation. The problem is the following. Suppose that we are given a binary function of a binary variable $f:\{0,1\}\rightarrow\{0,1\}$. Thus $f(0)$ can be either 0 or 1, and $f(1)$ likewise can be either 0 or 1, giving altogether four possibilities. However, suppose that we are not interested in the particular values of the function at 0 and 1, but we need to know whether the function is constant [i.e., $f(0)=f(1)$], or varying [i.e., $f(0)\neq f(1)$]. Deutsch poses the following task: by computing f only once, determine whether it is constant or varying. This kind of problem is generally referred to as a *promise algorithm*, because one property out of a certain number of properties is initially promised to hold, and our task is to determine computationally which one holds (see also Deutsch and Josza, 1992; Bernstein and Vazirani, 1993; and Simon, 1994) for other similar types of promise algorithms).

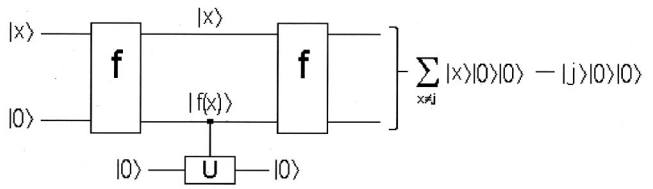

FIG. 9. A network that implements a phase-flip operation given the black-box computing the function $f(x)$. The unitary transformation U implements $|0\rangle \rightarrow -|0\rangle$ conditionally on the value of $f(x)$.

First of all, classically finding out in one step whether a function is constant or varying is clearly impossible. We need to compute $f(0)$ and then compute $f(1)$ in order to compare them. There is no way out of this double evaluation. Quantum mechanically, however, there is a simple method for performing this task by computing f only once. Two qubits are needed for the computation. In reality only one qubit is really needed, but the second qubit is there to implement the necessary transformation. We can imagine that the first qubit is the input to the quantum computer whose internal (hardware) part is represented by the second qubit. The computer itself will implement the following transformation on the two qubits (we perform this fully quantum mechanically, i.e., we are now not using "classical" devices such as beamsplitters):

$$|x\rangle|y\rangle \rightarrow |x\rangle|y \oplus f(x)\rangle, \tag{81}$$

where x is the input qubit and y is the hardware, as depicted in Fig. 9. Note that this transformation is reversible and thus there is a unitary transformation to implement it (but we shall not pay any attention to that at the moment, as we are interested here only in the basic principle). Note also that f has been used only once. The trick is to prepare the input in such a state that we make use of quantum superpositions. Let us have at the input

$$|x\rangle|y\rangle = (|0\rangle + |1\rangle)(|0\rangle - |1\rangle), \tag{82}$$

where $|x\rangle$ is the actual input and $|y\rangle$ is part of the computer hardware. Thus, before the transformation is implemented, the state of the computer is an equal superposition of all four basis states, which we obtain by simply expanding the state in Eq. (82),

$$|\Psi_{\text{in}}\rangle = |00\rangle - |01\rangle + |10\rangle - |11\rangle.$$

Note that there are negative phase factors before the

second and fourth terms. When this state now undergoes the transformation in Eq. (81), we have the following output state:

$$|\Psi_{\text{out}}\rangle = |0f(0)\rangle - |0\overline{f(0)}\rangle + |1f(1)\rangle - |1\overline{f(1)}\rangle$$
$$= |0\rangle(|f(0)\rangle - |\overline{f(0)}\rangle) + |1\rangle(|f(1)\rangle - |\overline{f(1)}\rangle),$$

where the overbar indicates the opposite of that value, so that, for example, $\overline{0}=1$. Now we see where the power of quantum computers is fully realized: each of the components in the superposition of $|\Psi_{in}\rangle$ underwent the same evolution of Eq. (81) "simultaneously," leading to the powerful "quantum parallelism" (Deutsch, 1985). This feature is true for quantum computation in general. Let us now look at the two possibilities.

(1) If f is constant then

$$|\Psi_{\text{out}}\rangle = (|0\rangle + |1\rangle)(|f(0)\rangle - |\overline{f(0)}\rangle).$$

(2) If f is varying then

$$|\Psi_{\text{out}}\rangle = (|0\rangle - |1\rangle)(|f(0)\rangle - |\overline{f(0)}\rangle).$$

Note that the output qubit (the first qubit) emerges in two different orthogonal states, depending on the type of f. These two states can be distinguished with 100% efficiency. This is easy to see if we first perform a Hadamard transformation on this qubit, leading to the state $|0\rangle$ if the function is constant and to the state $|1\rangle$ if the function is varying. Now a single projective measurement in 0,1 basis determines the type of the function. Therefore, unlike their classical counterparts, quantum computers can solve Deutsch's problem.

Let us now rephrase this in terms of phase shifts to emphasize its underlying identity with the above-mentioned Mach-Zender interferometer. The transformation of the two registers is the following:

$$|x\rangle|-\rangle \Rightarrow e^{i\pi f(x)}|x\rangle|-\rangle,$$

where $x=0,1$ and $|-\rangle = |0\rangle - |1\rangle$. Thus the first qubit is like a photon in the interferometer, receiving a conditional phase shift depending on its state (0 or 1). It is left to the reader to show that this transformation is formally identical to the above analysis. The second qubit is there just to implement the phase shift quantum mechanically. It should be emphasized that this quantum computation, although extremely simple, contains all the main features of successful quantum algorithms: it can be shown that all quantum computations are just more complicated variations of Deutsch's problem (Cleve *et al.*, 1997). We shall use the introduction of a phase shift as a basic element of a quantum computer and relate this to the notion of distinguishability and relative entropy.

Note one important aspect: the input could also be of the form $|-\rangle|-\rangle$. A constant function would then lead to the state $|-\rangle|-\rangle$ and a varying function would lead to $|+\rangle|-\rangle$. So the $|+\rangle$ and $|-\rangle$ are equally good as input states of the first qubit and both lead to quantum speedup. Their equal mixture, on the other hand, is not. This means that the output would be an equal mixture $|+\rangle\langle+|+|-\rangle\langle-|$ no matter whether $f(0)=f(1)$ or $f(0)\neq f(1)$, i.e., the two possibilities would be indistinguishable. Thus for the quantum algorithm to work well, we need the first register to be highly correlated to the two different types of functions. If the output state of the first qubit ρ_1 indicates that we have a constant function and ρ_2 that we have a varying function, then the efficiency of Deutsch's algorithm depends on how well we can distinguish the two states ρ_1 and ρ_2. This is given by the Holevo bound,

$$H = S(\rho) - \frac{1}{2}[S(\rho_1) + S(\rho_2)],$$

where $\rho = 1/2(\rho_1 + \rho_2)$. Thus, if $\rho_1 = \rho_2$, then $H=0$ and the quantum algorithm has no speedup over the classical one. At the other extreme, if ρ_1 and ρ_2 are pure and orthogonal, then $H=1$ and the computation gives the right result in one step. In between these two extremes lie all other computations with varying degrees of efficiency as quantified by the Holevo bound. Note that these are purely classical correlations and that there is no entanglement between the first and the second qubit. In fact, the Holevo bound is the same as the formula I suggested for classical correlations in the previous section. The key to understanding the efficiency of Deutsch's algorithm is therefore through the mixedness of the first register. If the initial state has the entropy of S_0, then the final Holevo bound is

$$S(\rho) - S_0.$$

So the more mixed the first qubit, the less efficient the computation. Note that the quantum mutual information between the first two qubits is zero throughout the entire computation (so there are neither classical nor quantum correlations between them).

B. Computation: Communication in time

Can we extend the above entropic analysis to other algorithms as well? The answer is yes, and this is exactly what I shall describe next (Bose, Rallan, and Vedral, 2000). To explain why this is so, I first need to introduce a few definitions and a communication model of quantum computation. We have two programmers, the sender and the receiver, and two registers, the memory (M) register and the computational (C) register. The sender prepares the memory register in a certain quantum state $|i\rangle_M$, which encodes the problem to be solved. For example, in the case of factorization (Shor, 1996), this register will store the number to be factored. In the case of a search (Grover, 1996), this register will store the state of the list to be searched. The number N of possible states $|i\rangle_M$ will, of course, be limited by the greatest number that the given computer could factor or the largest list that it could search. The receiver then prepares the computational register in some initial state ρ_C^0. Both the sender and the receiver feed the registers (prepared by them) to the quantum computer. The quantum computer implements the following general transformation on the registers:

$$(|i\rangle\langle i|)_M \otimes \rho_C^0 \rightarrow (|i\rangle\langle i|)_M \otimes U_i \rho_C^0 U_i^\dagger . \quad (83)$$

The resulting state $\rho_C(i) = U_i \rho_C^0 U_i^\dagger$ of the computational register contains the answer to the computation and is measured by the receiver. As the quantum computation should work for any $|i\rangle_M$, it should also work for any mixture $\Sigma_i^N p_i (|i\rangle\langle i|)_M$, where p_i are probabilities. For the sender to use the above computation as a communication protocol, he has to prepare any one of the states $|i\rangle_M$ with an *a priori* probability p_i. The entire input ensemble is thus $\Sigma_i^N p_i (|i\rangle\langle i|)_M \otimes \rho_C^0$. Due to the quantum computation, this becomes

$$\sum_i^N p_i (|i\rangle\langle i|)_M \otimes \rho_C^0 \rightarrow \sum_i^N p_i (|i\rangle\langle i|)_M \otimes \rho_C(i). \quad (84)$$

Whereas, before the quantum computation, the two registers were completely uncorrelated (mutual information is zero), at the end the mutual information becomes

$$I_{MC} := S(\rho_M) + S(\rho_C) - S(\rho_{MC}) = S(\rho_C) - \sum_i^N p_i S[\rho_C(i)], \quad (85)$$

where ρ_M and ρ_C are the reduced density operators for the two registers, ρ_{MC} is the density operator of the entire $M+C$ system, and $S(\rho) = -\mathrm{Tr}\rho \log \rho$ is the von Neumann entropy (for conventional reasons I shall use $\log_2$ in all calculations). Notice that the value of the mutual information (i.e., correlations) is equal to the Holevo bound $H = S(\rho_C) - \Sigma_i^N p_i S[\rho_C(i)]$ for the classical capacity of a quantum communication channel (Holevo, 1973). Note also that $\rho_C = \Sigma_i^N p_i \rho_C(i)$. This tells us how much information the receiver can obtain about the choice $|i\rangle_M$ made by the sender by measuring the computational register. The maximum value of H is obtained when the states $\rho_C(i)$ are pure and orthogonal. Moreover, the sender conveys the maximum information when all the message states have equal *a priori* probability (which also maximizes the channel capacity). In that case the mutual information (channel capacity) at the end of the computation is $\log N$. Thus the communication capacity I_{MC} [given by Eq. (85)] gives an index of the efficiency of a quantum computation. A necessary goal of a quantum computation is to achieve the maximum possible communication capacity consistent with given initial states of the quantum computer. We cannot give a sufficiency criterion from our general approach, as this depends on the specifics of an algorithm. If one breaks down the general unitary transformation U_i of a quantum algorithm into a number of successive unitary blocks, then the maximum capacity may be achieved only after a number of applications of the block. In each of the smaller unitary blocks, the mutual information between the M and the C registers (i.e., the communication capacity) increases by a certain amount. When its total value reaches the maximum possible value consistent with a given initial state of the quantum computer, the computation is regarded as being complete.

C. Black-box complexity

Any general quantum algorithm has to have a certain number of queries into the memory register (Bennett *et al.*, 1997; Beals *et al.*, 1998; Ambainis, 2000). This is necessitated by the fact that the transformation on the computational register has to depend on the problem at hand, encoded in $|i\rangle_M$. These queries can be considered to be implemented by a black box into which the states of both the memory and the computational registers are fed. The number of such queries needed in a certain quantum algorithm gives the black-box complexity of that algorithm (Bennett *et al.*, 1997; Beals *et al.*, 1998; Ambainis, 2000) and is a lower bound on the complexity of the whole algorithm. The black-box approach is a simplification for looking at the complexity of an algorithm. A black box allows us to perform a certain computation without having its exact details. It is possible that physical implementations of a particular black box may prove to be difficult. Therefore, when we estimate the complexity of an algorithm by counting the number of applications of a black box, we have to bear in mind that there might an additional complexity component arising in physical implementation.

In general we have a function $f:\{0,1\}^n \rightarrow \{0,1\}$ (so the function maps n-bit values to either 0 or 1). Quantum algorithms, such as a database search, can be expressed in this form (in the case of a database search, all the values of f are 0 apart from one that is equal to 1; the task is to find this value). The black box is assumed to be able to perform the transformation $|x\rangle|y\rangle \rightarrow |x\rangle|f(x) \oplus y\rangle$, just as in Deutsch's algorithm. We have the freedom to represent this black-box transformation as a phase flip which is equivalent in power (up to a constant factor as seen in Fig. 9),

$$|x\rangle|y\rangle \rightarrow (-1)^{f(x)\oplus y}|x\rangle|y\rangle.$$

Recently, Ambainis (2000) showed in a very elegant paper that if the memory register was prepared initially in the superposition $\Sigma_i^N |i\rangle_M$, then, in a search algorithm, $O(\sqrt{N})$ queries would be needed to completely entangle it with the computational register. This gives a lower bound on the number of queries in a search algorithm. In a manner analogous to his, I shall calculate the change in mutual information between the memory and the computational registers [from Eq. (85)] in one query step. The number of queries needed to increase the mutual information to $\log N$ (for perfect communication between the sender and the receiver) is then a lower bound on the complexity of the algorithm.

D. Database search

Any search algorithm, whether quantum or classical, regardless of its explicit form, will have to find a match for the state $|i\rangle_M$ of the M register among the states $|j\rangle_C$ of the C register and associate a marker to the state that matches (here $|j\rangle_C$ is a complete orthonormal basis for the C register). The most general way of doing such a

query in the quantum case is the black-box unitary transformation (Ambainis, 2000),

$$U_B|i\rangle_M|j\rangle_C=(-1)^{\delta_{ij}}|i\rangle_M|j\rangle_C. \tag{86}$$

Any other unitary transformation performing a query matching the states of the M and the C registers could be constructed from the above type of query. Note that the black box is able to recognize whether a value in the C register is the same as the solution, but is unable to explicitly provide that solution for us. For example, imagine that Socrates goes to visit the all-knowing ancient Greek oracle (black box), who is able to answer with only "yes" or "no." Suppose further that Socrates wants to know who is the wisest person in the world. He would then have to ask something like "Is Plato the wisest person in the world?" and would not be able to ask directly "Who is the wisest person in the world?" This "yes-no" approach is typical of any black-box analysis. The advantage of using this black box quantum mechanically is that we can query all the individual elements of the superposition simultaneously. Although we can identify the solution in one step quantum mechanically, further computations are required to amplify the right solution so that the subsequent measurement is more likely to reveal it.

I would like to put a bound on the change of the mutual information in one such black-box step. Let the memory states $|i\rangle_M$ be available to the sender with equal *a priori* probability so that the communication capacity is a maximum. His initial ensemble is then $1/N\,\Sigma_i^N(|i\rangle\langle i|)_M$. Let the receiver prepare the C register in an initial pure state ψ^0 [in fact, the power of quantum computation stems from the ability of the receiver to prepare pure-state superpositions of form $(1/N)\,\Sigma_j^N|j\rangle_C$]. This is an equal-weight superposition of all $|j\rangle_C$ as there is no *a priori* information about the right $|j\rangle_C$. This can be done by performing a Hadamard transformation on each qubit of the C register. In general, there will be many black-box steps on the initial ensemble before a perfect correlation is set up between the M and the C registers. After the kth black-box step, let the state of the system be

$$\rho^k=\frac{1}{N}\sum_i^N\,(|i\rangle\langle i|)_M\otimes[|\psi^k(i)\rangle\langle\psi^k(i)|]_C, \tag{87}$$

where

$$|\psi^k(i)\rangle_C=\sum_j\,\alpha_{ij}^k|j\rangle_C. \tag{88}$$

The $(k+1)$th black-box step changes this state to $\rho^{k+1}=(1/N)\,\Sigma_i^N(|i\rangle\langle i|)_M\otimes[|\psi^{k+1}(i)\rangle\langle\psi^{k+1}(i)|]_C$ with

$$|\psi^{(k+1)}(i)\rangle=\sum_{i,j}^N\,\alpha_{ij}^k(-1)^{\delta_{ij}}|j\rangle_C. \tag{89}$$

Thus we only have to evaluate the difference of mutual information between the M and the C register for the states. This difference of mutual information [when computed from Eq. (85)] can be shown to be the difference $|S(\rho_C^{k+1})-S(\rho_C^k)|$ (Henderson and Vedral, 2000). This quantity is bounded from above by (Fannes, 1999)

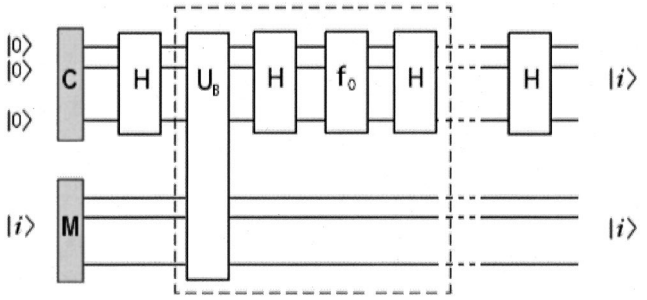

FIG. 10. The circuit for Grover's algorithm. C is the computational register and M is the memory register. U_B is the black-box query transformation, H is a Hadamard transformation on every qubit of the C register, and f_0 is a phase flip in front of the $|00\cdots0\rangle_C$. The block consisting of H, U_B, H, and f_0 is repeated a number of times.

$$|S(\rho_C^{k+1})-S(\rho_C^k)|\leqslant d_B(\rho_C^k,\rho_C^{k+1})\log N - d_B(\rho_C^k,\rho_C^{k+1})\log d_B(\rho_C^k,\rho_C^{k+1}), \tag{90}$$

where $d_B(\sigma,\rho)=\sqrt{1-F^2(\sigma,\rho)}$ is the Bures metric (Bures, 1969; Uhlmann, 1976, 1986) and $F(\sigma,\rho)=\mathrm{Tr}(\sqrt{\rho\sigma}\sqrt{\rho})^{1/2}$ is the fidelity. Using methods similar to those of Ambainis (2000), it can be shown that $F(\rho_C^0,\rho_C^1)\geqslant(N-2)/N$ from which it follows that the change in the first step is

$$|S(\rho_C^0)-S(\rho_C^1)|\leqslant\frac{3}{\sqrt{N}}\log N. \tag{91}$$

The change $|S(\rho_C^k)-S(\rho_C^{k+1})|$ in the subsequent steps has to be less than or equal to the change in the first step. This is because the Bures metric does not increase under general completely positive maps (which is what the query represents when we trace out the M register). Any other operations performed only on the C register in between two queries can only reduce the mutual information between the C and the M registers. This means that at least $O(\sqrt{N})$ steps are needed to produce full correlations (maximum mutual information of value $\log N$) between the two registers. This gives the black-box lower bound on the complexity of any quantum search algorithm. Of course, we know that there also exists an algorithm achieving this bound due to Grover (1996), and this has been proven to be optimal (Bennett, Bernstein, *et al.*, 1996; Zalka, 1999; Ambainis, 2000). However, the proof presented here is the most general, as it holds even when any type of completely positive map is allowed between queries [in Zalka (1999) a heuristic argument was made for the optimality of Grover's algorithm under general operations]. Grover's algorithm has also been implemented experimentally (Chuang *et al.*, 1998; Jones, Mosca, and Hansen, 1998).

I now use Grover's algorithm to show how the mutual information varies with time in a quantum search. The

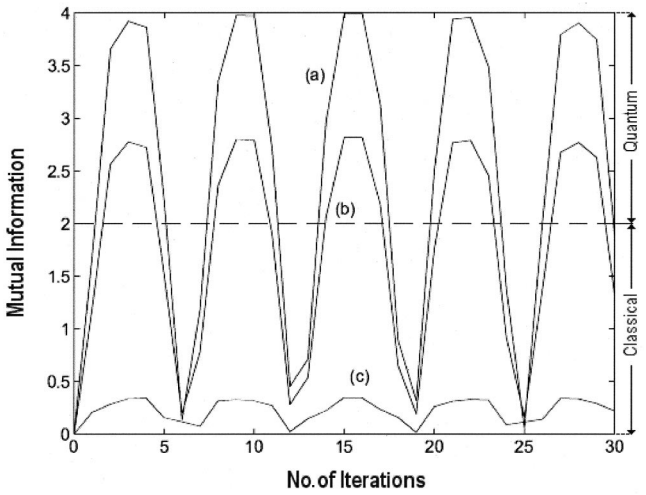

FIG. 11. Dependence of the mutual information between the M and the C registers as a function of the number of times the block in Grover's algorithm is iterated for various values of initial mixedness of the C register. Each qubit of the C register is initially in the state $p|0\rangle\langle 0|+(1-p)|1\rangle\langle 1|$; (a) $p=1$; (b) $p=0.95$; (c) $p=0.7$. The (a) and (b) computations achieve higher mutual information than classically allowed in the order of root N steps, while (c) does not.

general sequence described by Cleve *et al.* (1997) for Grover's algorithm will be used here. The algorithm consists of repeated blocks, each consisting of a Hadamard transform on each qubit of the C register, followed by a U_B (our black-box transformation), followed by another Hadamard transform on each qubit of the C register, and finally a phase flip f_0 of the $|00\cdots 0\rangle_C$ state of the C register (see Fig. 10). This block can then be repeated as many times as is necessary to bring the mutual information to its maximum value of $\log N$, which, as shown in Eq. (91), is $O(\sqrt{N})$. Note that the only transformation correlating the M and C registers is the black-box transformation U_B and that all the other transformations are done only on the C register and therefore do not change the mutual information between the two registers. In Fig. 8 I have plotted the variation of mutual information between the M and the C registers (i.e., the communication capacity of the quantum computation) with the number of iterations of the block in Grover's algorithm. It can be seen that the mutual information oscillates with the number of iterations. Figure 11 is plotted for a four-qubit computational register that can search a database of 16 entries. It reveals that the period is roughly 6, which means that the number of steps needed to achieve maximum mutual information is roughly 3. This is well above our bound for the minimum number of steps, which is 4/3 in this case.

The fact that the mutual information oscillates periodically (or more precisely, quasiperiodically) follows from the quantum Poincaré recurrence theorem (Hogg and Huberman, 1983), which states that if the system has a discrete spectrum and is "driven" by a periodic potential (as in Grover's case, where we repeat the same operation time and again), then its wave function $\psi(t)$ will undergo a quasiperiodic motion, i.e., for any $\epsilon>0$, there exists a relatively dense set $\{T_\epsilon\}$ such that

$$\|\psi(t+T_\epsilon)-\psi(t)\|<\epsilon$$

for all time t and for each T_ϵ in the set. This is exactly the behavior seen in Fig. 11. The distance between the two states $|\psi\rangle=\Sigma_i a_i|i\rangle$ and $|\phi\rangle=\Sigma_j b_j|j\rangle$ is defined in the usual way:

$$\|\psi-\phi\|:=\sum_i |a_i-b_i|.$$

The three graphs (a), (b), and (c) in Fig. 11 are for different values of initial mixedness of the C register. We find that the mutual information fails to rise to the maximum value of $\log N$ when the state of the computational register is mixed. Our formalism thus allows us to calculate the performance of a quantum computation as a function of the mixedness (quantified by the von Neumann entropy) of the computational register. We can put a bound on the entropy of the second register, after which the quantum search becomes as inefficient as the classical search. If the initial entropy $S(\rho_C^0)$ of the C register exceeds $\frac{1}{2}\log N$, then the change in mutual information between the M and the C registers in the course of the entire quantum computation would be at most $\log\sqrt{N}$. This can be achieved by a classical database search in $\sqrt{N}$ steps, so there is no advantage in using quantum evolution when the initial state is too mixed. Note that our condition

$$S(\rho_C^0)\geqslant\frac{1}{2}\log N$$

is only a sufficient and *not* a necessary condition for a quantum enhancement in efficiency.

I also point out that the states of the M register need not be a mixture, but could be an arbitrary superposition of states $|i\rangle_M$ [such a state was used by Ambainis (2000) in his argument]. All the above arguments still hold in that case, and the M and the C registers become quantum-mechanically entangled and not just classically correlated. Thus our analysis implies that any quantum computation is mathematically identical to a measurement process (Everett, 1973). The system being measured is the M register and the apparatus is the C register of the quantum computer. As time progresses the apparatus (register C) becomes more and more correlated (or entangled) to the system (register M). This means that the states of register C become more and more distinguishable, which allows us to extract more information about the M register by measuring the C register. The analysis in the last paragraph, in which I showed the limitations on the efficiency of quantum computation imposed by the mixedness of the C register, also applies to the efficiency of a quantum measurement when the apparatus is in a mixed state. Mixedness of an apparatus, to the best of our knowledge, has never been considered in the analysis of quantum measurement. In general practice, any apparatus, however macroscopic, is considered to be in a pure quantum state

before the measurement. Our approach highlighting the formal analogy between measurement and computation offers a way to analyze measurement in a much more general context.

Finally, I should like to discuss what would happen if we decided to change the nature of the black box. Suppose that, instead of being able to recognize the right solution, the black box is much more powerful and can determine whether the individual bit values coincide with the bit values of the solution. So, for all k,

$$|i_0 i_1 \cdots i_k \cdots i_n\rangle|j_0 i_j \cdots j_k \cdots j_n\rangle \rightarrow (-1)^{\delta_{i_k j_k}}|i_0 i_1 \cdots i_k \cdots i_n\rangle|j_0 i_j \cdots j_k \cdots j_n\rangle, \quad (92)$$

where $i=i_0 i_1 \cdots i_k \cdots i_n$ and $j=j_0 i_j \cdots j_k \cdots j_n$ are the binary representations of i and j, respectively. It can then easily be checked that this gate has the power to correlate the C and the M register by the amount of $\log 2$. Therefore the search algorithm would take $\log N$ steps (instead of $\sqrt{N}$), i.e., it would be polynomial instead of exponential. There is, of course, a hidden complexity here, which is in the construction of the new black box from the original black box. It can be shown that this requires an exponential increase in time (or space, which can always be traded for time), and this then compensates for the exponential decrease in the number of applications of the new black box. In fact, this new black box would be equivalent to the ancient Greek oracle's being able to answer the question posed by Socrates: "Who is the wisest person in the world?"

Can we use entropic measures of the above form to quantify the complexity of other quantum algorithms? The answer is unclear at present. The only algorithm that currently achieves an exponential speedup over its classical counterpart, Shor's factorization algorithm (Shor, 1996), cannot be usefully rephrased in terms of black-box operations (more precisely, it is rather trivial, as it requires only one black-box operation). However, this does not prevent us from deriving fundamental bounds on information storage and the speed of its processing based on the uncertainty principle. In the last subsection, I show the ultimate limits of processing power no matter what model of computation is used, so long as it uses quantum systems (particles or fields alike).

E. Quantum computation and quantum measurement

I now show that quantum computation is formally identical to a quantum measurement as described by von Neumann (1955). The analysis will be performed in the most general continuous case. Suppose that we have a system S (described by a continuous variable x) and an apparatus A (described by a continuous variable y interacting via a Hamiltonian $H=xp$), where p is the momentum of A (we shall assume that $\hbar=1$). Suppose in addition that the initial state of the total system is

$$|\Psi(0)\rangle = \int_x \phi(x)|x\rangle dx \otimes \eta(y)|y\rangle dy$$

in an uncorrelated state. The action of the above Hamiltonian then transforms the state into an entangled state. In order to calculate this transformation it will be beneficial to introduce the (continuous) Fourier transform

$$F_y : |y\rangle \rightarrow \int e^{-iyp}|p\rangle dp,$$

which takes us from the position space of A into the momentum space of A. This is important because we know the effect of the Hamiltonian in the momentum basis. Now, the action of the unitary transformation generated by H is

$$\begin{aligned}|\Psi(t)\rangle &= e^{-ixpt}|\Psi(0)\rangle \\ &= F_y e^{-ixpt} F_y |\Psi(0)\rangle \\ &= \int_x \int_y \phi(x)\,\eta(y-xy)|x\rangle|y\rangle dx dy,\end{aligned}$$

and we see that S and A are now correlated in x and y. This means that by measuring A we can obtain some information about the state of S. The mutual information $I_{AS}=H(x)+H(y)-H(x,y)$ can be shown to satisfy (Everett, 1973)

$$I_{AS} \geqslant \ln t,$$

i.e., it is growing at a rate faster than the logarithm of time passage during the measurement. This gives us a lower bound to exactly how quickly correlations can be established between the system and the apparatus. This is analogous to the way in which I derived the upper bound on the efficiency of quantum search algorithms in Sec. V.

Let us now calculate in greater detail the effect of the measurement Hamiltonian. We define

$$\xi(p) := F_y\{\eta(y)\}.$$

The evolution then proceeds as follows:

$$\begin{aligned}|\Psi(t)\rangle &= e^{-xpt}\int_x \phi(x)|x\rangle dx \otimes \eta(y)|y\rangle dy \\ &= e^{-xpt}\int_x \phi(x)\int_p \left\{\int_y \eta(y)e^{-iyp}dy\right\}|x\rangle|p\rangle dx dp \\ &= e^{-xpt}\int_x\int_p \phi(x)\xi(p)|x\rangle|p\rangle dx dp \\ &= \int_x \phi(x)\int_y \left\{\int_p \xi(p)e^{-ixpt}e^{iyp}dp\right\}|x\rangle|y\rangle dx dy \\ &= \int_x\int_y \phi(x)\,\eta(y-xy)|x\rangle|y\rangle dx dy.\end{aligned}$$

This result has the same formal structure as the quantum algorithms presented earlier: a Fourier transform followed by a conditional phase shift and then followed by another Fourier transform (cf. Deutsch's and Grover's algorithms). Therefore we can see that how efficiently we can measure something is the same as how efficiently

we can compute, both of which depend on how quickly we can establish correlations.

F. Ultimate limits of computation: The Bekenstein bound

Given a computer enclosed in a sphere of radius R and having available the total amount of energy E, what is the amount of information that it can store and how quickly can this information be processed? The Holevo bound gives us the ultimate answer. The amount of information that can be written into this volume is bounded from the above by the entropy, i.e., the number of distinguishable states that this volume can support. I shall now use a simple, informal argument to obtain this ultimate bound (Tipler, 1994), but the rigorous derivation can be found in Bekenstein (1981). The bound on energy implies a bound on momentum, and the total number of states in the phase space is

$$N=\frac{PR}{\Delta P\Delta R}\leqslant\frac{PR}{\hbar},$$

where the inequality follows from the Heisenberg uncertainty relation $\Delta P\Delta R\geqslant\hbar$, which limits the size of the smallest volume in the phase space to $\hbar$ in each of the three spatial directions. From relativity we have that for any particle $p\leqslant E/c$, so that

$$I\leqslant\ln N\leqslant N\leqslant\frac{E}{c}\frac{R}{\hbar}\leqslant\frac{ER}{\hbar c},$$

which is known as the Bekenstein bound. In reality this inequality will most likely be a huge overestimate, but it is important to know that no matter how we encode information we cannot perform better than is given by our most accurate present theory—quantum mechanics. As an example consider the nucleus of a hydrogen atom. According to the above result it can encode about 100 bits of information (I assumed that $E=mc^2$ and that $R=10^{-15}$ m). At present, NMR quantum computation achieves "only" one bit per nucleus (and not per nucleon)—spin up and spin down being the two states.

From the Bekenstein bound we can derive a bound on the efficiency of information processing. Again my derivation will be loose, and a much more careful calculation confirms what I shall present (Bekenstein, 1984). All the bits in the volume V cannot be processed faster than it takes light to travel across the volume $V=4/3\pi R^3$, which is $2R/c$. This gives

$$\frac{dI}{dt}\leqslant\frac{E}{2\hbar}.$$

Again a hydrogen nucleus can process 10^{24} bits per second, which is also in sharp contrast with NMR quantum computation where a NOT gate takes roughly a few milliseconds, leading to a maximum of 10^3 bits per second.

The Bekenstein bound shows that there is a potentially great number of underused degrees of freedom in any physical system. This provides hope that quantum computation will be an experimentally realizable goal. At present, there are a number of different practical implementations of quantum computation, but none of them can store and manipulate more than ten qubits at a time [five was the largest number (Vandersypen, 2000) that had been manipulated in a genuine quantum computation process at the time this review was finished in the summer of 2000]. The above calculation, however, does not take into account the environmental influence on computation nor the experimental precision. I have touched not at all on the practical possibility of building a quantum computer. This is partly for reasons of space, partly because it would spoil the flow of exposition, and partly because there are already a number of excellent reviews of this subject (Steane, 1997; Cory *et al.*, 2000). It is generally acknowledged that the difficulties in building a quantum computer are only of a practical nature and there are no fundamental limits that prohibit such a device. I hope that this section offers convincing arguments that building a quantum computer is very much a worthwhile adventure, from both the technological and the fundamental perspective. In any case we see that there is a great deal of currently unused potential in physical systems for storing and encoding information. As our level of technology improves we shall find more and more ways of getting close to the Bekenstein bound.

VI. CONCLUSIONS

We have seen how the distinguishability of different physical states is at the heart of information processing, which we quantified using the relative entropy. The relative entropy told us about the possibility of confusing two probability distributions, or, in the quantum case, two density matrices. We have seen that relative entropy never increases under any general quantum evolution, meaning that states can become only less distinguishable as time progresses. The most important consequence of this was shown to be the Holevo bound, which is the bound on the capacity for classical communication using quantum states. This basically told us that n qubits cannot store more than n classical bits of information. While this appears to be a severe limitation on quantum information processing, with the aid of dense coding quantum communication is in some sense more efficient than its classical counterpart. Dense coding involves the use of entangled states, and I therefore showed how the quantum relative entropy can be used to quantify entanglement. Moreover, I used the Holevo bound to put limits on the efficiency of quantum computation by treating it as a communication protocol. Quantum algorithms were shown to be considerably more efficient for some problems than classical algorithms. In particular, I have shown in a new way that the quantum database search has a square-root enhancement in efficiency over the classical database search. The efficiency of quantum computation stems from the tradeoff between two opposite effects: on the one hand, superpositions allow us to compute in parallel, while on the other hand, the Holevo bound limits the amount of information we can extract from a quantum state. I also emphasized links between black-box quantum computation and quantum measurement and I showed that there is a fundamental limit to

deleting information, leading to Landauer's principle that one bit erased increases the environment information by $k_B \ln 2$.

With every new physical theory comes a new understanding of the world we live in. Through Newtonian physics we understood the universe as a clockwork mechanism. With the subsequent development of thermodynamics the universe became a big Carnot engine, slowly evolving towards its final equilibrium state after which no useful work could be obtained—the heat death. At present we see the universe as an information-processing machine—a computer. Limits to the amount of information it can contain and process are given by the most accurate theory we have, quantum mechanics, giving rise to quantum information theory.

If there is a single moral to be drawn from the relationship between information and physics it is that, as we dig deeper into the fundamental laws of physics, we also push back the boundaries of information processing. It will not be surprising if all the results presented in this review are superseded by higher-level generalizations of which they become approximations in the same way in which classical information theory today approximates quantum information theory.

ACKNOWLEDGMENTS

I would like to thank S. Bose, D. Deutsch, D. P. DiVincenzo, M. J. Donald, A. Ekert, P. Hayden, L. Henderson, J. A. Jones, E. Kashefi, P. L. Knight, G. Lindblad, M. Murao, M. Ozawa, M. B. Plenio, L. Rallan, B. Schumacher, and G. Vidal for many stimulating discussions on the subject of quantum information. In particular, I thank L. Rallan for a very thorough reading of this manuscript and for helping me to draw some of the figures. Work for this review has been supported by the Knight Trust and Overseas Research Scheme Award, Elsag-Bailey spa, the Hewlett-Packard Company, Engineering and Physical Sciences Research Council, and the European Union project EQUIP (contract IST-1999-11053).

REFERENCES

Ambainis, A., 2000, in *Proceedings of the 32nd Annual ACM Symposium on the Theory of Computing*, May 2000 (Association for Computing Machinery, New York), p. 636.
Araki, H., and E. H. Lieb, 1970, Commun. Math. Phys. **18**, 160.
Barnum, H., M. A. Neilsen, and B. Schumacher, 1998, Phys. Rev. A **57**, 4153.
Beals, R., H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf, 1998, *Proceedings of the 39th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA), p. 352; also quant-ph/9802049.
Bekenstein, J. D., 1981, Phys. Rev. D **23**, 287.
Bekenstein, J. D., 1984, Phys. Rev. D **30**, 1669.
Bell, J., 1987, *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University, Cambridge).
Bennett, C. H., 1988, IBM J. Res. Dev. **32**, 16.
Bennett, C. H., E. Bernstein, G. Brassard, and U. Vazirani, 1997, SIAM J. Comput. **26**, 1510.
Bennett, C. H., H. J. Bernstein, S. Popescu, and B. Schumacher, 1996, Phys. Rev. A **53**, 2046.
Bennett, C. H., G. Brassard, C. Crepeau, R. Jozsa, A. Peres, and W. K. Wootters, 1993, Phys. Rev. Lett. **70**, 1895.
Bennett, C. H., D. P. DiVincenzo, C. A. Fuchs, T. Mor, E. Rains, P. W. Shor, J. A. Smolin, and W. K. Wootters, 1999, Phys. Rev. A **59**, 1070.
Bennett, C. H., D. P. DiVincenzo, T. Mor, P. W. Shor, J. A. Smolin, and B. M. Terhal, 1999, Phys. Rev. Lett. **82**, 5385.
Bennett, C. H., D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, 1996, Phys. Rev. A **54**, 3824.
Bennett, C. H., and S. Wiesner, 1992, Phys. Rev. Lett. **69**, 2881.
Bernstein, E., and U. Vazirani, 1993, in *Proceedings of the 25th Annual ACM Symposium on the Theory of Computing* (ACM, New York), p. 11.
Bhatia, R., 1997, *Matrix Analysis* (Springer-Verlag, Berlin).
Boschi, D., S. Branca, F. DeMartini, L. Hardy, and S. Popescu, 1998, Phys. Rev. Lett. **80**, 1121.
Bose, S., M. B. Plenio, and V. Vedral, 2000, J. Mod. Opt. **47**, 291.
Bose, S., L. Rallan, and V. Vedral, 2000, Phys. Rev. Lett. **85**, 5448.
Bouwmeester, D., J. W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, 1997, Nature (London) **390**, 575.
Bowen, G., 2001, Phys. Rev. A **63**, 022302.
Brillouin, L., 1956, *Science and Information Theory* (Academic, New York).
Bures, D., 1969, Trans. Am. Math. Soc. **135**, 199.
Caves, C. M., and P. D. Drummond, 1994, Rev. Mod. Phys. **66**, 481.
Choi, M. D., 1975, Linear Algebr. Appl. **10**, 285.
Chuang, I. L., N. Gershenfeld, and M. Kubinec, 1998, Phys. Rev. Lett. **80**, 3408.
Clauser, J. F., M. A. Horne, A. Shimony, and R. A. Holt, 1969, Phys. Rev. Lett. **23**, 880.
Cleve, R., A. Ekert, C. Macchiavello, and M. Mosca, 1997, Philos. Trans. R. Soc. London, Ser. A **454**, 339.
Cory, D. G., R. Laflamme, E. Knill, L. Viola, T. F. Havel, N. Boulant, G. Boutis, E. Fortunato, S. Lloyd, R. Martinez, C. Negrevergne, M. Pravia, Y. Sharf, G. Teklemariam, Y. S. Weinstein, and W. H. Zurek, 2000, Fortschr. Phys. **48**, 875.
Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory* (Wiley, New York).
Csiszár, I., and J. Körner, 1981, *Coding Theorems for Discrete Memoryless Systems* (Academic, New York).
Davies, E. B., 1976, *Quantum Theory of Open Systems* (Academic, London).
Deutsch, D., 1985, Proc. R. Soc. London, Ser. A **400**, 97.
Deutsch, D., 1989, Proc. R. Soc. London, Ser. A **425**, 73.
Deutsch, D., 1998, *The Fabric of Reality* (Viking-Penguin, London).
Deutsch, D., and R. Jozsa, 1992, Proc. R. Soc. London, Ser. A **439**, 553.
DiVincenzo, D. P., C. A. Fuchs, H. Mabuchi, J. A. Smolin, A. Thapliyal, and A. Uhlmann, 1999, *Quantum Computing and Communications, Lecture Notes in Computer Science* (Springer, Berlin), Vol. 1505, p. 247.
DiVincenzo, D. P., P. W. Shor, J. A. Smolin, B. M. Terhal, and A. V. Thapliyal, 2000, Phys. Rev. A **61**, 062312.
Donald, M. J., 1986, Commun. Math. Phys. **105**, 13.

Donald, M. J., 1987, Math. Proc. Cambridge Philos. Soc. **101**, 363.
Donald, M. J., 1992, Found. Phys. **22**, 1111.
Donald, M. J., and M. Horodecki, 1999, Phys. Lett. A **264**, 257.
Einstein, A., B. Podolsky, and N. Rosen, 1935, Phys. Rev. **47**, 777.
Eisert, J., T. Felbinger, P. Papadopoulos, M. B. Plenio, and M. Wilkens, 2000, Phys. Rev. Lett. **84**, 1611.
Ekert, A., and R. Jozsa, 1996, Rev. Mod. Phys. **68**, 733.
Everett, H., III, 1957, Rev. Mod. Phys. **29**, 454.
Everett, H., III, 1973, in *The Many-Worlds Interpretation of Quantum Mechanics*, edited by B. S. DeWitt and N. Graham (Princeton University, Princeton, NJ), p. 3.
Fannes, M., 1973, Commun. Math. Phys. **31**, 291.
Feynman, R. P., 1996, *Feynmann Lectures on Computation*, edited by A. J. G. Hey and R. W. Allen (Addison-Wesley, Reading, MA).
Fuchs, C. A., 1996, *Distinguishability and Accessible Information in Quantum Theory*, Ph.D. thesis (University of New Mexico); also e-print quant-ph/9601020.
Furusawa, A., J. L. Sorensen, S. L. Braunstein, C. A. Fuchs, H. J. Kimble, and E. S. Polzik, 1998, Science **282**, 706.
Garey, M., and D. Johnson, 1979, *Computers and Intractability: A Guide to the Theory of NP-completeness* (Freeman, San Francisco).
Garisto, R., and L. Hardy, 1999, Phys. Rev. A **60**, 827.
Gisin, N., 1996, Phys. Lett. A **210**, 151, and references therein.
Gordon, J. P., 1964, in *Quantum Electronics and Coherent Light*, Proceedings International School of Physics "Enrico Fermi," Course 31, edited by P. A. Miles (Academic, New York), p. 156.
Grover, L. K., 1996, *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing* (ACM, New York), p. 212; also e-print quant-ph/9605043.
Hausladen, P., R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters, 1996, Phys. Rev. A **54**, 1869.
Hayashi, M., 2001, J. Phys. A **34**, 3413.
Hayden, P. M., M. Horodecki, and B. M. Terhal, 2001, J. Phys. A **34**, 3413.
Henderson, L., and V. Vedral, 2000, Phys. Rev. Lett. **84**, 2263.
Hiai, F., and D. Petz, 1991, Commun. Math. Phys. **143**, 99.
Hogg, T., and B. A. Huberman, 1983, Phys. Rev. A **28**, 22.
Holevo, A. S., 1973, Probl. Peredachi Inf. **9**, 3. [Probl. Inf. Transm. **9**, 177 (1973)].
Holevo, A. S., 1982, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam).
Holevo, A. S., 1998, IEEE Trans. Inf. Theory **44**, 269.
Horodecki, M., 1998, Phys. Rev. A **57**, 3364.
Horodecki, M., P. Horodecki, and R. Horodecki, 1996, Phys. Lett. A **223**, 1.
Horodecki, M., P. Horodecki, and R. Horodecki, 1998, Phys. Rev. Lett. **80**, 5239.
Horodecki, M., P. Horodecki, and R. Horodecki, 2000, Phys. Rev. Lett. **84**, 2014.
Horodecki, M., P. Horodecki, and R. Horodecki, 2001, Phys. Lett. A **283**, 1.
Ingarden, R. S., 1976, Rep. Math. Phys. **10**, 43.
Ingarden, R. S., A. Kossakowski, and M. Ohya, 1997, *Information Dynamics and Open Systems—Classical and Quantum Approach* (Kluwer Academic, Dordrecht).
Jaynes, E. T., and F. W. Cummings, 1963, Proc. IEEE **51**, 89.
Jones, J. A., and M. Mosca, 1998, J. Chem. Phys. **109**, 1648.
Jones, J. A., M. Mosca, and R. H. Hansen, 1998, Nature (London) **393**, 344.
Jozsa, R., and B. Schumacher, 1994, J. Mod. Opt. **41**, 2343.
King, C., and M. B. Ruskai, 2001, J. Math. Phys. **42**, 87.
Kolmogorov, A. N., 1950, *Foundations of the Probability Theory* (Chelsea, New York).
Kraus, B., J. I. Cirac, S. Karnas, and M. Lewenstein, 2000, Phys. Rev. A **61**, 062302.
Kraus, K., 1983, *States, Effects and Operations: Fundamental Notions of Quantum Theory*, Lecture Notes in Physics No. 180 (Springer, Berlin).
Kullback, S., and R. A. Leibler, 1951, Ann. Math. Stat. **22**, 79.
Landauer, R., 1961, IBM J. Res. Dev. **5**, 183.
Lebedev, D. S., and L. B. Levitin, 1963, Sov. Phys. Dokl. **8**, 377.
Lesniewski, A., and M. B. Ruskai, 1999, J. Mod. Phys. **40**, 5702.
Lewenstein, M., D. Bruss, J. I. Cirac, B. Kraus, M. Kus, J. Samsonowicz, A. Sanpera, and R. Tarrach, 2000, J. Mod. Opt. **47**, 2481.
Lieb, E. H., and M. B. Ruskai, 1973, J. Math. Phys. **14**, 1938.
Lindblad, G., 1974, Commun. Math. Phys. **39**, 111.
Lindblad, G., 1975, Commun. Math. Phys. **40**, 147.
Lo, H., and S. Popescu, 1999, Phys. Rev. Lett. **83**, 1459.
Mackey, G. W., 1963, *Mathematical Foundations of Quantum Mechanics* (W. A. Benjamin, New York).
Nielsen, M., 1999, Phys. Rev. A **61**, 064301.
Ohya, M., and D. Petz, 1993, *Quantum Entropy and Its Use*, Texts and Monographs in Physics (Springer-Verlag, Berlin).
Ozawa, M., 1984, J. Math. Phys. **25**, 79.
Papadimitriou, C. H., 1995, *Computational Complexity* (Addison-Wesley, New York).
Partovi, M. H., 1989, Phys. Lett. A **137**, 445, and references therein.
Penrose, O., 1973, *Foundations of Statistical Mechanics* (Oxford University, Oxford).
Peres, A., 1993, *Quantum Theory: Concepts and Methods* (Kluwer Academic, Dordrecht).
Peres, A., 1996, Phys. Rev. Lett. **77**, 1413.
Plenio, M. B., and V. Vedral, 1998, Contemp. Phys. **38**, 431.
Popescu, S., and D. Rohrlich, 1997, Phys. Rev. A **56**, R3319.
Rains, E. M., 1999a, Phys. Rev. A **60**, 173.
Rains, E. M., 1999b, Phys. Rev. A **60**, 179.
Redhead, M., 1987, *Incompleteness, Nonlocality and Realism* (Clarendon, Oxford).
Reed, M., and B. Simon, 1980, *Methods of Modern Mathematical Physics-Functional Analysis* (Academic, New York).
Sanov, I. N., 1957, Mat. Sb. **42**, 11.
Schmidt, E., 1907, Math. Ann. **63**, 433.
Schrödinger, E., 1935, Naturwissenschaften **23**, 807, 823, 844.
Schumacher, B., 1995, Phys. Rev. A **51**, 2738.
Schumacher, B., 1996, Phys. Rev. A **54**, 2614.
Schumacher, B., and M. D. Westmoreland, 1997, Phys. Rev. A **56**, 131.
Schumacher, B. W., and M. D. Westmoreland, 2000, e-print quant-ph/0004045.
Shannon, C. E., and W. Weaver, 1949, *The Mathematical Theory of Communication* (University of Illinois, Urbana, IL).
Shor, P. W., 1996, in *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science*, November 1994, edited by S. Goldwasser (IEEE Computer Society, Los Alamitos, CA), p. 124.

Simon, D. S., 1994, *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science*, edited by S. Goldwasser (IEEE Computer Society, Los Alamitos, CA), p. 16.
Steane, A., 1997, Appl. Phys. B: Lasers Opt. **B64**, 623.
Stern, T. E., 1960, IEEE Trans. Inf. Theory **6**, 435.
Tipler, F. J., 1994, *The Physics of Immortality* (Bantam Doubleday Dell, New York).
Toffoli, T., 1981, Math. Syst. Theory **14**, 13.
Tolman, R. C., 1938, *The Principles of Statistical Mechanics* (Oxford University, Oxford).
Uhlmann, A., 1976, Rep. Math. Phys. **9**, 273.
Uhlmann, A., 1986, Rep. Math. Phys. **24**, 229.
Umegaki, H., 1962, Kodaikanal Math. Sem. Rep. **14**, 59.
Vandersypen, L. M. K., M. Steffen, G. Breyta, C. S. Yannoni, R. Cleve, and I. L. Chuang, 2000, Phys. Rev. Lett. **85**, 5452.
Vedral, V., 2000, Proc. R. Soc. London, Ser. A **456**, 969.
Vedral, V., and M. B. Plenio, 1998, Phys. Rev. A **57**, 1619.
Vedral, V., M. B. Plenio, K. Jacobs, and P. L. Knight, 1997, Phys. Rev. A **56**, 4452.
Vedral, V., M. B. Plenio, M. A. Rippin, and P. L. Knight, 1997, Phys. Rev. Lett. **78**, 2275.
Vedral, V., M. A. Rippin, and M. B. Plenio, 1997, J. Mod. Opt. **44**, 2185.
Vidal, G., 2000, J. Mod. Opt. **47**, 355.
Vollbrecht, K. G. H., and R. F. Werner, 2001, Phys. Rev. A **64**, 062307.
von Neumann, J., 1955, *Mathematical Foundations of Quantum Mechanics*, translated from the German ed. by R. T. Beyer (Princeton University, Princeton).
Wehrl, A., 1978, Rev. Mod. Phys. **50**, 221.
Werner, R. F., 1989, Phys. Rev. A **40**, 4277.
Wootters, W. K., 1998, Phys. Rev. Lett. **80**, 2245.
Wootters, W. K., and W. H. Zurek, 1982, Nature (London) **299**, 802.
Yamamoto, Y., and H. A. Haus, 1986, Rev. Mod. Phys. **58**, 1001.
Yuen, H. P., and M. Ozawa, 1993, Phys. Rev. Lett. **70**, 363.
Zalka, C., 1999, Phys. Rev. A **60**, 2746.