Classical information capacity of a quantum channel

Paul Hausladen,[1] Richard Jozsa,[2] Benjamin Schumacher,[3] Michael Westmoreland,[4] and William K. Wootters[5]

[1]*Department of Physics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6396*

[2]*School of Mathematics and Statistics, University of Plymouth, Plymouth, Devon PL4 8AA, England*

[3]*Department of Physics and Astronomy, Kenyon College, Gambier, Ohio 43022*

[4]*Department of Mathematics, Denison University, Granville, Ohio 43023*

[5]*Department of Physics, Williams College, Williamstown, Massachusetts 01267*

(Received 26 June 1995; revised manuscript received 14 May 1996)

We consider the transmission of classical information over a quantum channel. The channel is defined by an ''alphabet'' of quantum states, e.g., certain photon polarizations, together with a specified set of probabilities with which these states must be sent. If the receiver is restricted to making separate measurements on the received ''letter'' states, then the Kholevo theorem implies that the amount of information transmitted per letter cannot be greater than the von Neumann entropy H of the letter ensemble. In fact the actual amount of transmitted information will usually be significantly less than H. We show, however, that if the sender uses a block coding scheme consisting of a choice of code words that respects the *a priori* probabilities of the letter states, and the receiver distinguishes whole words rather than individual letters, then the information transmitted per letter can be made arbitrarily close to H and never exceeds H. This provides a precise information-theoretic interpretation of von Neumann entropy in quantum mechanics. We apply this result to ''superdense'' coding, and we consider its extension to noisy channels. [S1050-2947(96)12209-5]

PACS number(s): 03.65.Bz, 89.70.+c

I. QUANTUM CHANNELS

Quantum information theory concerns the transmission and manipulation of information stored in systems that must be treated quantum mechanically. As in classical information theory, one of the most basic questions in quantum information theory is this: How efficiently can one transmit information using a given set of resources? In contrast to the classical case, however, quantum theory presents two very different forms of the question, depending on the nature of the information to be conveyed. On the one hand, one can try to convey quantum states themselves: the sender has a quantum system in an unknown state and wants the receiver to end up with a similar system in the same state. A coding theorem for this case has recently been proved [1,2]. On the other hand, one might want to *use* quantum states to convey *classical* information, that is, information that can be expressed as a sequence of zeros and ones. The situation is particularly interesting if the quantum states one is using are not all orthogonal to each other, in which case they cannot be distinguished from each other perfectly by the receiver. This is the problem we consider here. As we will see, this problem leads to a new information-theoretic interpretation of the von Neumann entropy of an ensemble of states.

Nonorthogonal quantum states might be used in a variety of contexts to transmit classical information. In studies of quantum cryptography, nonorthogonal signals are used intentionally in order to avoid eavesdropping [3]. Moreover, in any transmission using signals at the quantum level, such as weak coherent pulses in an optical fiber, any ambiguity between signals may be more a matter of nonorthogonality (e.g., overlapping pulses) than classical noise. In what follows we will often imagine the signals to be nonorthogonal photon polarization states, but our analysis applies to all quantum systems.

Suppose that a sender, Alice, wishes to transmit classical information to a receiver, Bob, using a communication channel that is quantum mechanical (for instance, the polarization of a photon). Alice will represent possible messages by preparing the channel in various quantum states. Bob will recover the information by subjecting the channel to a measurement. As noted above, however, unless the signal states used by Alice are orthogonal, no measurement will allow Bob to distinguish perfectly between them. Thus, Bob's ability to recover Alice's message without error will be limited by the quantum mechanics of the channel. Indeed no ''decoding observable'' will be sufficient to recover the entire information content of the message in the quantum signal source. It is therefore more appropriate to consider the *accessible information*, the maximum amount of information about the message that can be recovered in a measurement performed on the system M that conveys that message. The proper measure of recovered information is the *mutual information*, which for a pair of random variables X and Y is defined to be

$$I(X:Y)=H(X)-H(X|Y). \tag{1}$$

Here H is the Shannon entropy, which is a function of the probabilities $p(x_i)$ of the possible values of X:

$$H(X)=-\sum_i p(x_i)\log_2 p(x_i). \tag{2}$$

(Here we interpret $p\log_2 p$ as taking the value zero when $p=0$.) $H(X|Y)$ is the expected entropy of X once one knows the value of Y. That is,

$$H(X|Y)=\sum_j p(y_j)\left[-\sum_i p(x_i|y_j)\log_2 p(x_i|y_j)\right]. \quad (3)$$

In classical information theory, the mutual information is the amount of information about X that is acquired by determining the value of Y. The crucial theorem from classical information theory that justifies our focus on $I(X:Y)$ is this: if a communication channel has mutual information $I(X:Y)$ between the input signal X and the received output Y, then by means of sufficiently redundant coding, that channel can be used to send up to, but no more than, $I(X:Y)$ binary digits per use of the channel with arbitrarily low probability of error. In our quantum context, if we denote by B the outcome of a measurement of an observable on M, the quantity $I(A:B)$ measures the information about the message source A that is acquired by measurement of the observable. It also, therefore, measures the number of binary digits that can be conveyed per signal when the receiver uses this observable.

A theorem stated by Gordon [4] and Levitin [5] and first proved by Kholevo [6] states that the amount of information accessible to Bob is limited by the entropy of the ensemble of signal states. That is, suppose Alice represents each message a (with *a priori* probability p_a) by a state ρ_a, which is in general a mixed state. Then for any observable that Bob chooses to measure, the mutual information $I(A:B)$ between Alice's input A and Bob's measurement outcome B is bounded by

$$I(A:B)\leqslant H(\rho)-\sum_a p_a H(\rho_a), \quad (4)$$

where $\rho=\Sigma_a p_a\rho_a$ (the average density matrix for the ensemble $\mathcal{E}$ of signals) and $H(\rho)=-\mathrm{Tr}\rho\ln\rho$ (the von Neumann entropy of the density matrix ρ). If the signal states ρ_a are all pure states, the second term on the right vanishes, and we can simply say

$$I(A:B)\leqslant H(\rho). \quad (5)$$

As Kholevo noted [7], there are situations in which $I(A:B)$ does not approach $H(\rho)$ very closely for any choice of Bob's decoding observable. Thus, though this theorem provides an *upper bound* on the amount of information accessible to Bob, this upper bound is not in general very strong [8,9].

One example of this was studied in detail by Peres and Wootters [10]. Suppose Alice sends a photon in one of three linear polarization states (called ''letter'' states) denoted $|a\rangle$, $|b\rangle$, and $|c\rangle$, which are separated by 120°. Each state is used equally often. This signal ensemble has a von Neumann entropy $H(\rho)$ of 1.000 bit, whereas Bob's optimal decoding observable yields a mutual information $I(A:B)$ of 0.585 bit. If two photons are used, there are nine possible signal states: $|aa\rangle$, $|ab\rangle$, etc. The von Neumann entropy is 2.000 bits and the optimal mutual information available to Bob is 1.170 bits.

However, suppose Alice sends two photons but only uses the three states $|aa\rangle$, $|bb\rangle$, and $|cc\rangle$. Note that the individual letter states are being used with their original (equal) probabilities in this restricted choice of two-photon states. Then the ensemble entropy $S(\rho)$ is only 1.585 bits, but the optimal mutual information is 1.369 bits, or about 0.685 bit per photon. In other words, by restricting her code to a subset of the possible code words, while still respecting the given letter frequencies, Alice can increase the distinguishability of the code words and increase the information conveyed per photon to Bob.

This example shows that it is sometimes possible to increase the accessible information per elementary signal by (a) using code words composed of several elementary signals, and (b) *deleting* some of the possible code words in the ensemble, while respecting the given elementary signal frequencies. The receiver then chooses a decoding observable optimized to distinguish among the code words actually used. Note that this observable will typically *not* be realizable as a set of separate measurements on the individual elementary signals; rather, it will be a joint measurement on the whole set of signals constituting a code word.

These considerations lead us to ask whether it is possible for Alice and Bob to use this strategy to approach the Kholevo bound. That is, given an *a priori* ensemble of pure-state signals with entropy $H(\rho)$, can Alice and Bob choose a set of code words respecting the original probabilities of the signals, together with a decoding observable, so that information is reliably transmitted at a rate approaching $H(\rho)$ per elementary signal? The answer is yes. Moreover, we will see that for no such code can the transmission rate exceed $H(\rho)$.

We now give a precise formulation of our main result. Suppose we are given an ensemble $\mathcal{E}$ of *letter states* $|a\rangle$ of an elementary quantum system (not necessarily a photon) with *a priori* probabilities p_a. The letter ensemble has a density matrix

$$\rho=\sum_a p_a|a\rangle\langle a|, \quad (6)$$

with von Neumann entropy $H(\rho)=-\mathrm{Tr}\rho\ln\rho$.

A *code* [''(N,l) code''] consists of two things: (i) a set of N code words $\{|s_i\rangle:i=1,\ldots,N\}$ where each code word is a sequence (i.e., product) of l letter states (but not all such sequences of letter states are code words), and (ii) an *a priori* probability p_{s_i} assigned to each code word. The *tolerance* τ of the code is defined by

$$\tau=\max_a|f_a-p_a|, \quad (7)$$

where f_a is the overall frequency of occurrence of the letter $|a\rangle$ among the Nl letters of all the code words, taking into account the *a priori* probabilities of the code words. That is,

$$f_a=\frac{1}{l}\sum_{i=1}^{N} p_{s_i}n_{a,s_i}, \quad (8)$$

where n_{a,s_i} is the number of occurrences of letter $|a\rangle$ in code word $|s_i\rangle$. A low tolerance code will use up the letters approximately with their given *a priori* probabilities p_a in the construction of the code words.

Then we may consider the information transmissible using the ensemble of code word states. The information per

letter is just the accessible information of the code word ensemble divided by l. We shall show the following:

Theorem: Let I_δ be the least upper bound on the information per letter transmissible with any code having tolerance $\leqslant \delta$. Then

$$\lim_{\delta\to 0} I_\delta = H(\rho).$$

This theorem gives a precise information-theoretic interpretation of von Neumann entropy in quantum mechanics. To put it in somewhat looser but more familiar terms, the theorem says that if Alice is required to use certain quantum states as signals with certain specified frequencies of occurrence, the number of binary digits she can convey to Bob per particle can be made arbitrarily close to, but not greater than, the von Neumann entropy of the ensemble of signals.

Before proving the theorem, it is interesting to compare our result to the channel capacity theorem for classical channels [15], which describes the reliable transmission of information through a noisy classical channel. For a given set of input letters, transmissible through a noisy channel (but with the *a priori* probabilities unspecified), the classical channel capacity C is defined to be the maximum attainable mutual information, where the maximum is taken over all possible choices of probability distribution of the input letters. Then the classical theorem states that C is the maximum possible rate at which information can be reliably transmitted through the noisy channel.

In our quantum theorem above, we have considered a fixed probability distribution p_a for the input letters $|a\rangle$, which allows us to characterize the von Neumann entropy of the letter ensemble. However, we may consider this distribution to be variable (for a fixed set of letter states) and *define* the capacity C of the quantum channel to be the maximum von Neumann entropy attainable from the given letter states by free choice of the *a priori* probabilities:

$$C := \max_{p_a} H(\rho). \tag{9}$$

Then according to our theorem, C is the maximum rate at which classical information may be transmitted through the quantum channel (the channel being characterized entirely by its allowed letter states). Below, in the course of the proof of the quantum theorem we will show that for any input distribution p_a and for any $\epsilon, \delta > 0$ there exists a code and a decoding observable such that the amount of information transmitted per letter is greater than $H(\rho) - \delta$ and the probability of error is less than ϵ. (In fact this will be achieved with code words having equal *a priori* probabilities.) Thus information may be reliably transmitted at rate C and no more, which is formally similar to the statement of the classical channel capacity theorem and justifies the definition (9) above, of quantum channel capacity in the present context.

Despite this similarity between classical channel capacity and the above notion of quantum channel capacity, there are also essential differences, in particular the origin of the conditional probabilities in Eq. (3). In the classical setting these probabilities are fixed, being characteristic of the noise in the channel. On receiving the (partially corrupted) signals, Bob simply records them and the issue of his measurement is trivial. In contrast in our quantum setting, letter states are transmitted to Bob without alteration; i.e., there is no ''noise.'' The problem now is that nonorthogonal quantum states are not perfectly distinguishable by any measurement. Bob has much freedom in his choice of decoding measurement and the conditional probabilities in Eq. (3) arise from the probabilistic nature of quantum measurements. Our notion of quantum channel capacity (9) corresponds to optimizing over choice of input distribution *and* of decoding observable, in contrast to classical channel capacity, which involves maximizing only over input distribution. Indeed in the quantum case, even if the input distribution is viewed as being fixed (as in our theorem above) the conditional probabilities in (1) are still variable as we need to optimize over all possible choices of the decoding observable. It is this requirement that necessitates the discussion in Sec. III.

The *proof* of the theorem presented here is similar in some respects to that of the classical channel capacity theorem: both rely on the concatenation of letter states to obtain code words and the pruning of the set of all possible code words to obtain the codes to be used in the channel. In the classical setting the aim of pruning is to increase redundancy whereas in the quantum setting, it is to increase *distinguishability* of the code word states. It should be emphasized that this concatenation and pruning do not result in a channel that differs from the original. In both situations the question is as follows: If we allow for repeated transmission of elementary letter states, what is the maximum rate at which information can be conveyed? Both proofs utilize a method of random coding to demonstrate the existence of a code with desired properties. This is merely a technique of proof and *not a method of signaling*. Considering averages over random codes is a means of bringing statistical arguments to bear upon the problem. It does not provide a means of constructing the required code.

In the next five sections we describe the machinery necessary to complete the proof, including typical subspaces of quantum ensembles, a sufficiently optimal choice of decoding observable, and the technique of ''random coding.'' The final sections discuss some of the corollaries and consequences of the main result.

II. TYPICAL SUBSPACES

Suppose we have the ensemble of letter states $|a\rangle$ in a Hilbert space $\mathcal{H}$ with probabilities p_a, as described above. Fix $\epsilon, \delta > 0$. We can concatenate letters into words of length l, and if l is long enough [1,2] then the Hilbert space $\mathcal{H}^l = \mathcal{H}\otimes\mathcal{H}\otimes\cdots\otimes\mathcal{H}$ (the lth-order tensor power of $\mathcal{H}$) for the words can be decomposed into two subspaces: a ''typical'' subspace Λ and the perpendicular subspace $\Lambda^\perp$, having the following properties: (a) Both Λ and $\Lambda^\perp$ are spanned by eigenstates of $\rho^l = \rho\otimes\rho\otimes\cdots\otimes\rho$ (the lth-order tensor power of ρ). (b) Almost all of the weight of the ensemble lies within Λ: $\mathrm{Tr}\Pi_\Lambda \rho^l \Pi_\Lambda > 1-\epsilon$ and $\mathrm{Tr}\Pi_{\Lambda^\perp}\rho^l\Pi_{\Lambda^\perp} < \epsilon$. (Here Π_Λ and $\Pi_{\Lambda^\perp}$ are the projections onto Λ and $\Lambda^\perp$.) (c) The eigenvalues λ_n of ρ^l for eigenstates in Λ fall within a ''narrow'' range:

$$2^{-l[H(\rho)+\delta]} < \lambda_n < 2^{-l[H(\rho)-\delta]}. \tag{10}$$

[Of course, $H(\rho^l) = lH(\rho)$.] (d) The number of dimensions in the typical subspace is bounded in the range

$$(1-\epsilon)2^{l[H(\rho)-\delta]} \leqslant \dim\Lambda \leqslant 2^{l[H(\rho)+\delta]}. \tag{11}$$

The typical subspace Λ is constructed as follows: The eigenvalues q_i of the one-letter density operator ρ form a ''probability distribution'' for the eigenstates of ρ, for which the classical (Shannon) entropy is just the von Neumann entropy $H(\rho)$. Eigenstates of ρ^l are sequences of ρ eigenstates. By the weak law of large numbers, we can find a set of ''typical'' eigenstates of ρ^l in which the frequencies of the ρ eigenstates are close to the ''probabilities'' q_i. Λ is the subspace spanned by these typical eigenstates.

Suppose we sum the squares of the eigenvalues of ρ^l, but restrict ourselves to the typical subspace. Then we get

$$\begin{aligned}\mathrm{Tr}\Pi_\Lambda(\rho^l)^2\Pi_\Lambda &< (\dim\Lambda)(2^{-l[H(\rho)-\delta]})^2 \\ &< 2^{-l[H(\rho)-3\delta]}.\end{aligned} \tag{12}$$

This inequality will be useful below in connecting the probability of error for a coding scheme to the entropy $H(\rho)$.

III. DECODING OBSERVABLE

Suppose that Alice is using a code with words long enough for the typical subspace to exist and that have the properties outlined above. Alice may not be using *all* of the possible code words. Denote an individual code word by $|s_i\rangle$.

In order to read Alice's messages, Bob will have to employ a decoding observable to determine which signal $|s_i\rangle$ is present. This decoding observable will in general be a ''positive operator'' measurement (or POM) [11,12]. Bob will want to choose his decoding observable so that he will deduce Alice's message with as small a probability of error as possible.

Bob's essential problem is to distinguish between a collection of vectors in the Hilbert space $\mathcal{H}^l$. Let $\{|\phi_k\rangle\}$ be a collection of such vectors (possibly not normalized). Consider the operator

$$\Phi = \sum_k |\phi_k\rangle\langle\phi_k|, \tag{13}$$

which is a positive operator whose support is the subspace spanned by the vectors $|\phi_k\rangle$. On this subspace $\Phi^{1/2}$ exists and is invertible, so we can form the vectors [13]

$$|\mu_k\rangle = \Phi^{-1/2}|\phi_k\rangle \tag{14}$$

corresponding to positive operators $|\mu_k\rangle\langle\mu_k|$. These positive operators are easily shown to be a resolution of the identity on this subspace:

$$\begin{aligned}\sum_k |\mu_k\rangle\langle\mu_k| &= \sum_k \Phi^{-1/2}|\phi_k\rangle\langle\phi_k|\Phi^{-1/2} \\ &= \Phi^{-1/2}\left(\sum_k |\phi_k\rangle\langle\phi_k|\right)\Phi^{-1/2} \\ &= \Phi^{-1/2}\Phi\Phi^{-1/2} = 1.\end{aligned} \tag{15}$$

The operators $|\mu_k\rangle\langle\mu_k|$, supplemented by a projection onto the subspace perpendicular to the span of $\{|\phi_k\rangle\}$, thus form the outcome operators of a POM.

The vectors $|\phi_k\rangle$ specify a particular POM that employs the outcome operators $|\mu_k\rangle\langle\mu_k|$. This is the POM that Bob chooses in order to distinguish among the vectors. This is a reasonable choice. If the vectors $|\phi_k\rangle$ are orthogonal and thus completely distinguishable, the resulting measurement does indeed distinguish them perfectly. (There is no known way of specifying the *best* observable in general, but this observable will be good enough for our purposes.)

The $|\mu_k\rangle$ vectors have another interesting and (for us) useful property. Let S_{jk} be the matrix of inner products among the $|\phi_k\rangle$ vectors:

$$S_{jk} = \langle\phi_j|\phi_k\rangle. \tag{16}$$

If there are N vectors, this is an $N\times N$ complex matrix with positive eigenvalues. The $|\mu_k\rangle$ vectors are related to the square root of this matrix by

$$(\sqrt{S})_{jk} = \langle\mu_j|\phi_k\rangle. \tag{17}$$

In fact, this property of the $|\mu_j\rangle$ vectors can be taken as an implicit definition for them.

To decode Alice's message, Bob will employ an observable to distinguish between her signal states $|s_i\rangle$. But we will find it more useful to suppose that he distinguishes between projections of the signal states into the typical subspace Λ—that is, between non-normalized vectors $|\sigma_i\rangle = \Pi_\Lambda|s_i\rangle$. To do this he will employ the ''square root'' measurement just described. Since the typical subspace contains ''almost all'' of the set of available code words (in the sense of the previous section), this refinement introduces negligible error, as we shall show. Thus, we define the matrix $S_{ij} = \langle\sigma_i|\sigma_j\rangle$, and construct the $|\mu_i\rangle$ vectors (which lie within Λ) so that

$$\langle\mu_i|s_j\rangle = \langle\mu_i|\sigma_j\rangle = (\sqrt{S})_{ij}. \tag{18}$$

Let us also define $n_i = S_{ii} = \langle\sigma_i|\sigma_i\rangle$, the norm of the projected code words. The $|\mu_k\rangle$ vectors, together with the projection onto the subspace perpendicular to all the vectors $|\sigma_i\rangle$, define Bob's POM.

IV. PROBABILITY OF ERROR

Alice's code will consist of N code words $|s_i\rangle$, each used with equal frequency. The information content of a single code word is therefore $\log_2 N$. Each code word is a sequence of l letters chosen from the set of possible letters. (For now, we will disregard the probabilities of those letters in the given ensemble.) Bob devises his decoding observable as described above.

Alice sends the signal $|s_i\rangle$ with probability $1/N$. Bob will correctly interpret the signal—that is, he will obtain the μ_i outcome in his decoding POM—with probability

$$p(\mu_i|s_i) = \mathrm{Tr}|\mu_i\rangle\langle\mu_i|s_i\rangle\langle s_i| = |\langle\mu_i|s_i\rangle|^2. \tag{19}$$

We note that $\langle\mu_i|s_i\rangle$ is real and non-negative. The average probability of error is thus

$$P_E = 1 - \sum_i \frac{1}{N} \langle \mu_i | s_i \rangle^2 = \frac{1}{N} \sum_i (1 - \langle \mu_i | s_i \rangle)(1 + \langle \mu_i | s_i \rangle)$$

$$\leq \frac{2}{N} \sum_i (1 - \langle \mu_i | \sigma_i \rangle). \tag{20}$$

In terms of the S_{ij} matrix, this is

$$P_E \leq \frac{2}{N} \sum_i [1 - (\sqrt{S})_{ii}]. \tag{21}$$

The square root function is bounded below by a parabola: for $x \geq 0$,

$$\sqrt{x} \geq \tfrac{3}{2} x - \tfrac{1}{2} x^2. \tag{22}$$

The matrix S is a matrix with non-negative eigenvalues, so this inequality may be applied to it:

$$\sqrt{S} \geq \tfrac{3}{2} S - \tfrac{1}{2} S^2. \tag{23}$$

This means that, for a complex N vector with components z_k,

$$\sum_{kl} z_k^* (\sqrt{S})_{kl} z_l \geq \frac{3}{2} \sum_{kl} z_k^* S_{kl} z_l - \frac{1}{2} \sum_{klj} z_k^* S_{kj} S_{jl} z_l. \tag{24}$$

For a given i, we can choose $z_i = 1$ and $z_k = 0$ for $k \neq i$. This yields

$$(\sqrt{S})_{ii} \geq \frac{3}{2} S_{ii} - \frac{1}{2} \sum_j S_{ij} S_{ji}$$

$$= \frac{3}{2} n_i - \frac{1}{2} n_i^2 - \frac{1}{2} \sum_{j \neq i} S_{ij} S_{ji}. \tag{25}$$

Therefore

$$P_E \leq \frac{2}{N} \sum_i \left(1 - \frac{3}{2} n_i + \frac{1}{2} n_i^2 + \frac{1}{2} \sum_{j \neq i} S_{ij} S_{ji} \right)$$

$$= \frac{2}{N} \sum_i \left((1 - n_i)(1 - n_i/2) + \frac{1}{2} \sum_{j \neq i} S_{ij} S_{ji} \right)$$

$$\leq \frac{2}{N} \sum_i \left((1 - n_i) + \frac{1}{2} \sum_{j \neq i} S_{ij} S_{ji} \right). \tag{26}$$

V. RANDOM CODES

In this section and the next we will prove our main result. We will show that Alice can choose N code words with N sufficiently large so that $\log_2 N$ is approximately $lH(\rho)$, such that Bob (using the scheme above) has probability of error P_E nearly equal to zero. Furthermore we will see that Alice's code can be chosen to have arbitrarily small tolerance τ as defined in (7). Finally we will show that an information rate of $H(\rho)$ bits per letter cannot be exceeded in the limit of vanishing tolerance.

To show the existence of such a code, we will in fact show that almost any code will do the job. That is, we will calculate the average probability of error for an ensemble of *random* codes of N words. We generate a random code in this way. Each of the N code words is a sequence of l letter states generated using the *a priori* probabilities for the letters. The probability that the ith code word $|s_i\rangle = |a_1 a_2 \cdots a_l\rangle$ is just $p(a_1)p(a_2)\cdots p(a_l)$. Each code word is generated independently of the others. We are in effect drawing N code words at random *with replacement* from the *a priori* ensemble.

Denote an average over random codes by $\langle\ \rangle_c$. First, we note that, for any particular code word $|s_i\rangle$,

$$\langle |s_i\rangle\langle s_i| \rangle_c = \rho^l. \tag{27}$$

Next we take the average of P_E over random codes.

$$\langle P_E \rangle_c \leq \frac{2}{N} \sum_i \left(1 - \langle n_i \rangle_c + \frac{1}{2} \sum_{j \neq i} \langle S_{ij} S_{ji} \rangle_c \right). \tag{28}$$

Each of the averages in this expression is straightforward to calculate. The average norm $\langle n_i \rangle_c$ of the ith projected code word is

$$\langle n_i \rangle_c = \langle \mathrm{Tr}(\Pi_\Lambda |s_i\rangle\langle s_i| \Pi_\Lambda) \rangle_c$$

$$= \mathrm{Tr}(\Pi_\Lambda \rho^l \Pi_\Lambda) \geq 1 - \epsilon. \tag{29}$$

For $j \neq i$, the code words $|s_j\rangle$ and $|s_i\rangle$ are independent, so that

$$\langle S_{ij} S_{ji} \rangle_c = \langle \langle s_i | \Pi_\Lambda | s_j \rangle \langle s_j | \Pi_\Lambda | s_i \rangle \rangle_c$$

$$= \langle \mathrm{Tr}(\Pi_\Lambda |s_i\rangle\langle s_i| |s_j\rangle\langle s_j| \Pi_\Lambda) \rangle_c$$

$$= \mathrm{Tr} \Pi_\Lambda (\rho^l)^2 \Pi_\Lambda. \tag{30}$$

(We have used the fact that Π_Λ commutes with ρ^l.)

Each term in the upper bound for $\langle P_E \rangle_c$ is independent of i, so that the sum over i yields a multiplicative factor of N. The j sum yields a factor of $N-1$. Thus

$$\langle P_E \rangle_c < 2\epsilon + N \mathrm{Tr} \Pi_\Lambda (\rho^l)^2 \Pi_\Lambda. \tag{31}$$

We use the properties of typical subspaces [Eq. (12) above] to obtain

$$\langle P_E \rangle_c < 2\epsilon + N 2^{-l[H(\rho) - 3\delta]}. \tag{32}$$

If the *average* probability of error is below this bound, then Alice and Bob will be able to find some particular code for which

$$P_E < 2\epsilon + N 2^{-l[H(\rho) - 3\delta]}. \tag{33}$$

If l is very large (perhaps much larger than we actually need to form the typical subspace Λ), Alice can make $N = 2^{l[H(\rho) - 4\delta]}$ and still have $P_E < 3\epsilon$. In this case, Alice encodes $H(\rho) - 4\delta$ bits per letter. This proves the existence of codes allowing transmission at an asymptotic rate of $H(\rho)$ bits per letter, with arbitrarily low error.

Remark: In fact, we can do even better: we can modify a code with a low *average* probability of error to give a code with a low *maximum* probability of error. Let us throw away

the worst half of the code words in the optimizing code. Since the average probability of error for this code is less than 3ϵ, we have

$$\frac{1}{2^{lH(\rho)}}\sum_i [1-p(\mu_i|s_i)]\leqslant 3\epsilon. \tag{34}$$

This implies that at least half the code words must have conditional probability of error of less than 6ϵ; otherwise, these code words would contribute at least 3ϵ to the sum. Thus, in the reduced code book we have $2^{lH(\rho)-1}$ code words. By throwing out half of the code words we have changed the rate from $H(\rho)$ to $H(\rho)-1/l$, a negligible difference for large l.

VI. LETTER FREQUENCIES AND CHANNEL CAPACITY

The argument in the previous section is nonconstructive. We have not explicitly constructed a code with low probability of error; we have merely shown that such a code must exist. Therefore we do not know much about its detailed properties. However, we now show that the code may be chosen to have arbitrarily small tolerance.

Note first that the generation of a random code consisting of N code words of length l amounts to an independent choice of Nl letters according to the *a priori* probability distribution. Since each code word is used equally often, the number of times a given letter appears in the list of N code words tells us its frequency of occurrence when the code is used by Alice and Bob. We can apply the weak law of large numbers to the set of codes as follows: if Nl is sufficiently large, the set of all random codes may be divided into two classes: (a) a set of ''typical'' codes, in which the letter frequencies approximate the *a priori* probabilities to within a fixed tolerance; and (b) a set of ''atypical'' codes, which are generated by random coding with small total probability. The ''atypical'' codes, having small total probability, contribute very little to the overall average probability of error $\langle P_E\rangle_c$ estimated above. Thus, $\langle P_E\rangle_c$ must also be very small even if Alice and Bob are restricted to using ''typical'' codes—every one of which has letter frequencies matching the *a priori* probabilities p_a to within any specified tolerance.

A different consideration arises if Alice and Bob are not required to use any particular letter frequencies but are free to adjust them as they please in order to maximize the information per letter conveyed by their channel. For that case, as discussed in Sec. I, we may define the *channel capacity* C of a quantum channel with a particular alphabet $\{|a\rangle\}$ to be

$$C=\max_{p_a} H(\rho), \tag{35}$$

where $\rho=\Sigma_a p_a|a\rangle\langle a|$. Our argument above implies that Alice may communicate with arbitrarily low probability of error, up to C bits per letter, to Bob using the letter states $\{|a\rangle\}$.

To complete the proof of the main result we now show that the information rate $H(\rho)$ bits per letter cannot be exceeded in the limit of vanishing tolerance. Consider any (N,l) code having (small) tolerance δ in which the code words $|s_i\rangle$ $i=1,\ldots,N$ have *a priori* probabilities p_{s_i} (not assumed to be equal). Let ρ_{code} and H_{code} denote the density matrix and von Neumann entropy of the code word ensemble.

Let $\mathcal{E}_1$ be the ensemble of letter states that appear as first letters in the code words; i.e., we look at the first letter of each of the N code words and note the frequency $f_a^{(1)}$ of occurrence of each letter $|a\rangle$, taking into account the *a priori* probabilities of the code words. Let ρ_1 and H_1 be the density matrix and entropy of $\mathcal{E}_1$. Thus

$$\rho_1=\sum_a f_a^{(1)}|a\rangle\langle a|. \tag{36}$$

Similarly define $\mathcal{E}_k$, ρ_k, and H_k for each position $k=1,\ldots,l$ in the code words.

Now each word is a product state of letters so that ρ_k is the reduced state of ρ_{code}, obtained by partial tracing over all letter positions except the kth position. Hence subadditivity of von Neumann entropy [16] gives

$$H_{\text{code}}\leqslant H_1+\cdots+H_l. \tag{37}$$

Equation (37) holds regardless of any tolerance constraints and already implies that the quantum channel capacity C, as defined in (9), is an upper bound on the information rate per letter. Indeed, since each $\mathcal{E}_k$ is an ensemble of letter states, it follows from the definition of C that $H_k\leqslant C$. From (37) we get

$$H_{\text{code}}/l\leqslant(H_1+\cdots+H_l)/l\leqslant C \tag{38}$$

and the Kholevo theorem applied to the code word ensemble gives that the transmitted information per letter is less than H_{code}/l. Thus an information rate of C bits per letter cannot be exceeded and the random coding argument in the previous section shows that this upper bound is tight.

Let us now incorporate tolerance constraints. Let us suppose that the tolerance δ is extremely small. This means that to high accuracy the average density matrix of all Nl letters in all the code words is ρ:

$$\rho=\frac{1}{l}\sum_{k=1}^{l}\rho_k. \tag{39}$$

Now concavity of von Neumann entropy [16] gives

$$H(\rho)\geqslant\frac{H_1+\cdots+H_l}{l}. \tag{40}$$

Then (40) and (37) give $H_{\text{code}}/l\leqslant H(\rho)$ so that by the Kholevo theorem

$$\text{(information)/(letter)}\leqslant\text{(entropy)/(letter)}=H_{\text{code}}/l\leqslant H(\rho), \tag{41}$$

as required.

Equation (39) is strictly true only if the tolerance is zero. For small nonzero tolerance (37) remains exact but (39) becomes

$$\frac{1}{l}\sum_{k=1}^{l}\rho_k=\rho^*, \tag{42}$$

where $|H(\rho^*)-H(\rho)|\to 0$ as $\delta\to 0$. Equation (41) becomes

$$(\text{information})/(\text{letter}) \leqslant H(\rho^*) \tag{43}$$

and our desired result appears in the limit as $\delta\to 0$. This completes the demonstration of our main theorem.

VII. SUPERDENSE CODING

We can apply our results to an interesting quantum communication scheme proposed by Bennett and Wiesner [14], which has been called ''superdense coding.'' Superdense coding makes use of the quantum entanglement between systems to enhance their information capacity.

The simplest example works as follows. Alice and Bob initially share a pair of two-state systems—e.g., a pair of spins—which are in an entangled state. Suppose that this state is one of the four orthogonal ''Bell states,'' given by

$$|\Psi^{\pm}\rangle=\frac{1}{\sqrt{2}}(|\uparrow_1\downarrow_2\rangle\pm|\downarrow_1\uparrow_2\rangle),$$
$$|\Phi^{\pm}\rangle=\frac{1}{\sqrt{2}}(|\uparrow_1\uparrow_2\rangle\pm|\downarrow_1\downarrow_2\rangle). \tag{44}$$

For definiteness, we imagine that the initial state is the singlet $|\Psi^-\rangle$.

Alice manipulates her own spin and then transmits it to Bob, who performs a measurement on both spins. Ordinarily, the transmission of a single spin could only communicate one bit of information to Bob (by a simple application of the Kholevo theorem). However, in this case the pre-established entanglement between the pair of spins will allow Alice to send *two* bits of information to Bob. This can happen because Alice can convert the initial state $|\Psi^-\rangle$ into any one of the four Bell states by a suitable unitary transformation on only one of the spins, and Bob can distinguish between the four orthogonal Bell states by a coherent measurement of the pair of spins. Alice's four-way choice encodes a two-bit message that is perfectly recoverable by Bob.[1]

We are interested in a more general situation. Alice and Bob work with N-state quantum systems instead of spins, and they may possess a considerable supply of them (so that they may use block coding of many independent messages). If Alice were to send messages to Bob by sending N-state quantum systems, she could send up to $\log_2 N$ bits per system. However, suppose that Alice and Bob share many pairs of systems, which are each in some entangled pure state (which may or may not be ''maximally entangled'' like the Bell states). What is the information capacity of these entangled systems for superdense coding?

We can write the initial state of one of the entangled pairs using the Schmidt decomposition:

$$|\Psi_0\rangle=\sum_{k=1}^{N}\sqrt{p_k}|a_k b_k\rangle, \tag{45}$$

where $|a_k\rangle$ is an orthonormal basis for Alice's system and $|b_k\rangle$ is a basis for Bob's system. The density matrix for Bob's system given by a partial trace over Alice's system has eigenvalues p_k. We will call the entropy H_E of that density matrix the *entropy of entanglement* of the system. H_E will be between zero (for a product state) and $\log_2 N$ (for a maximally entangled state).

Alice can perform a unitary transformation on her system, after which she delivers it to Bob. We might imagine her performing different transformations with different *a priori* probabilities, leading to an ensemble of states for the pair of systems that Bob measures. Our theorem establishes that, by judicious coding (and choice of Bob's decoding observable), Alice may convey an amount of information up to the entropy of this ensemble. How big may this entropy be—i.e., what is the information capacity of this scheme?

It is easy to see that the entropy can be *no larger* than $H_E+\log_2 N$, since Alice's manipulations of her system do not affect the density matrix for Bob's system. The total entropy for the pair of systems cannot be greater than the entropy of Bob's system (which is H_E) plus the largest possible entropy for Alice's system (which is $\log_2 N$). It can also be shown that a particular ensemble of transformations can make the overall entropy equal to $H_E+\log_2 N$. One such ensemble of transformations would include all permutations of the Schmidt basis states $|a_k\rangle$, rotations of the relative phases of these states, and combinations of the two.

We can therefore conclude that the channel capacity of the superdense coding scheme is $H_E+\log_2 N$. This is a sensible result. If the pair of systems is initially in a product state, $H_E=0$ and so Alice can only send $\log_2 N$ bits per system, as expected. If the pair of systems is maximally entangled, then the capacity is $2\log_2 N$, exactly twice as great.

VIII. NOISY CHANNELS

So far we have assumed that when Alice sends a letter state $|a\rangle$, the state arrives at Bob's end unchanged. In many practical applications, however, the channel will introduce noise and the signal will arrive in some mixed state ρ_a. In that case, it is as if Alice were using an ensemble of mixed states to send her message rather than an ensemble of pure states. Our theorem does not apply to ensembles of mixed states, but it suggests the following conjecture concerning this case.

Let Alice be given an ensemble $\mathcal{E}$ of letter states ρ_a with *a priori* probabilities p_a, and let ρ be the the density matrix of the whole ensemble: $\rho=\Sigma_a p_a\rho_a$. We conjecture that the amount of information Alice can convey per letter can be made arbitrarily close to the quantity χ defined by

$$\chi=H(\rho)-\sum_a p_a H(\rho_a). \tag{46}$$

Note that χ is the upper bound that appears in the general form (4) of the Kholevo theorem.

To argue for this conjecture, it is helpful to consider two different ensembles: the ensemble $\mathcal{E}$ of mixed states that Alice is given, and the ensemble $\mathcal{E}'$ consisting of all the eigenstates $|a,j\rangle$ of the density matrices ρ_a. The *a priori* probability of the state $|a,j\rangle$ in $\mathcal{E}'$ is $p_a q_{aj}$, where q_{aj} is the

[1]The apparent doubling of the information capacity in the presence of entanglement motivates the name ''superdense coding.''

eigenvalue of ρ_a corresponding to $|a,j\rangle$. In other words, $\mathcal{E}'$ is a refinement of $\mathcal{E}$, in which each mixed state is replaced by its pure eigenstates.

Consider now a random code constructed from the original mixed-state ensemble $\mathcal{E}$. This code corresponds to a specific code constructed from $\mathcal{E}'$: in place of each $\mathcal{E}$ code word, substitute the set of *all* corresponding $\mathcal{E}'$ code words (in which each ρ_a is replaced by one of its eigenstates), with probabilities determined by the eigenvalues. There is no physical difference between these two codes as regards the set of possible transmissions. What is different in the $\mathcal{E}'$ code is that Alice knows which pure state she sends.

Let us suppose for now that our main theorem applies to $\mathcal{E}'$ codes constructed in this way. That is, if Alice had the ability to know which pure signal she was sending, then she could convey up to $H(\rho)$ bits per letter using a typical $\mathcal{E}'$ code constructed as above. Included in this information, however, is the information that Bob would obtain about which specific eigenstates were used. In actuality, Alice knows only the mixed $\mathcal{E}$ code word, so that the information Bob obtains about which eigenstates were used is irrelevant to Alice's message. Now, the amount of information Bob could have obtained per letter about these eigenstates is the average entropy of the mixed signal states ρ_a, that is, $\Sigma_a p_a H(\rho_a)$. Thus, from the additivity of information it follows that the amount of information Bob actually obtains about which $\mathcal{E}$ code word Alice sent can be made arbitrarily close to

$$H(\rho)-\sum_a p_a H(\rho_a), \tag{47}$$

which is what we wanted to show.

Unfortunately our theorem as it stands does not apply to the $\mathcal{E}'$ codes because of the lack of strict independence among the code words. However, it is plausible that a modified argument could account for this case and thereby prove the conjecture.

IX. CONCLUSION

We have shown that the von Neumann entropy $H(\rho)$ of an ensemble of pure quantum states is equal to the capacity of a quantum channel to transmit classical information where the quantum channel transmits the states with their given *a priori* probabilities. This conclusion holds in spite of the fact that for all nonorthogonal ensembles, the amount of information one can obtain by a measurement on a *single* system is strictly less than $H(\rho)$ [6]. One achieves the increased capacity by having the receiver discriminate among whole code words rather than trying to distinguish the individual signal states.

Considering the importance of entropy in other contexts, it is satisfying that in this communication problem the entropy turns out to be the actual channel capacity, and not merely an upper bound.

ACKNOWLEDGMENT

R.J. acknowledges support of the Leverhulme Trust and the Royal Society, London.

[1] B. Schumacher, Phys. Rev. A **51**, 2738 (1995).

[2] R. Jozsa and B. Schumacher, J. Mod. Opt. **41**, 2343 (1994).

[3] S. Wiesner, Sigact News **15**, 78 (1983); C. H. Bennett and G. Brassard, in *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing, Bangalore, India* (IEEE, New York, 1984), pp. 175–179; A. K. Ekert, Phys. Rev. Lett. **67**, 661 (1991); C. H. Bennett, G. Brassard, and N. D. Mermin, *ibid.* **68**, 557 (1992); C. H. Bennett, *ibid.* **68**, 3121 (1992); A. K. Ekert, J. G. Rarity, P. R. Tapster, and G. M. Palma, *ibid.* **69**, 1293 (1992).

[4] J. P. Gordon, in *Quantum Electronics and Coherent Light, Proceedings of the International School of Physics ''Enrico Fermi,'' Course XXXI*, edited by P. A. Miles (Academic, New York, 1964), pp. 156–181.

[5] L. B. Levitin, in *Proceedings of the Fourth All-Union Conference on Information and Coding Theory, Sec. II, Tashkent, 1969*, translated by A. Bezinger and S. L. Braunstein in *Quantum Communication and Measurement*, edited by R. Hudson, V. P. Belavkin, and O. Hirota (Plenum, New York, 1995).

[6] A. S. Kholevo, Probl. Peredachi Inf. **9**, 3 (1973) [Probl. Inf. Transm. (USSR) **9**, 177 (1973)].

[7] A. S. Kholevo, Probl. Peredachi Inf. **9**, 110 (1973) [Probl. Inf. Transm. (USSR) **9** (2), 31 (1973)].

[8] C. A. Fuchs and C. M. Caves, Phys. Rev. Lett. **73**, 3047 (1994).

[9] R. Jozsa, D. Robb, and W. K. Wootters, Phys. Rev. A **49**, 668 (1994).

[10] A. Peres and W. K. Wootters, Phys. Rev. Lett. **66**, 1119 (1991).

[11] C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976), pp. 74–83.

[12] A. Peres, Found. Phys. **20**, 1441 (1990).

[13] L. P. Hughston, R. Jozsa, and W. K. Wootters, Phys. Lett. A **183**, 14 (1993).

[14] C. H. Bennett and S. J. Wiesner, Phys. Rev. Lett. **69**, 2881 (1992).

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[16] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).