Sending classical information via noisy quantum channels

Benjamin Schumacher[1] and Michael D. Westmoreland[2]
[1]Department of Physics, Kenyon College, Gambier, Ohio 43022
[2]Department of Mathematical Sciences, Denison University, Granville, Ohio 43023

(Received 12 February 1997)

This paper extends previous results about the classical information capacity of a noiseless quantum-mechanical communication channel to situations in which the final signal states are mixed states, that is, to channels with noise. [S1050-2947(97)02007-6]

PACS number(s): 03.65.Bz, 42.50.Dv, 89.70.+c

I. INTRODUCTION

Suppose Alice wishes to convey classical information to Bob by using a quantum system Q as a communication channel. Alice prepares the channel in one of various quantum states W_x with *a priori* probabilities p_x. Bob makes a measurement on the system Q, and from its result he tries to infer which state Alice prepared. A theorem stated by Gordon [1] and Levitin [2], proved by Kholevo [3], gives an upper bound to the amount of information that Bob can obtain about Alice's signal. If $W=\Sigma_x p_x W_x$ is the density operator describing the ensemble of Alice's signals, then the mutual information $H(X\!:\!Y)$ between Alice's input X and Bob's output Y is bounded by

$$H(X\!:\!Y) \leqslant H(W) - \sum_x p_x H(W_x), \tag{1}$$

where $H(W) = -\mathrm{Tr} W \log_2 W$, the von Neumann entropy of the density operator W. The upper bound in Eq. (1) is in general a weak one, in that Bob may not be able to choose an observable that gives him an amount of information near the upper bound [4].

Recently, Hausladen *et al.* [5] showed that, if Alice's signal states W_x are pure states, then it is possible to approach the Kholevo bound $H(W)$ for an appropriate choice of Alice's code and Bob's decoding observable. This is done by (i) employing long strings of signals to send many independent messages together, (ii) ''pruning'' the set of strings used as codewords so that the codewords are sufficiently distinguishable, and (iii) choosing a suitable decoding observable that acts on entire strings of signals. For large enough L, codewords of L ''letters'' may be used to transmit up to $LH(W)$ bits of information [thus $H(W)$ bits per letter] with arbitrarily low probability of error.

This naturally suggests a generalization, which was presented in [5] as a conjecture. Suppose that Alice employs signal states W_x that are *mixed* states. Then can Alice and Bob find a choice of code and decoding observable so that the general Kholevo bound [Eq. (1)] can be approached arbitrarily closely? In this paper, we show that the answer to this question is ''yes.'' That is, we prove the following result.

Theorem. Suppose we have letter states W_x with *a priori* probabilities p_x and let

$$\chi = H(W) - \sum_x p_x H(W_x).$$

Fix $\epsilon, \delta > 0$. Then for sufficiently large L, there exist a code (whose codewords are strings of L letters) and a decoding observable such that the information carried per letter is at least $\chi - \delta$ and the probability of error $P_E < \epsilon$.

As in [5], we employ an average over randomly generated codes to establish the existence of a satisfactory code. (If the average probability of error is small for an ensemble of codes, the ensemble must contain specific codes with small probability of error.) We also use a similar prescription for Bob's decoding observable. The chief refinement in the proof presented here is the enforcement of stronger ''typicality'' conditions on various quantities associated with the channel.

The mixed states W_x may be thought of as the outputs of a *noisy* quantum channel. Thus our main result will enable us to draw conclusions about the classical information capacity of a noisy quantum channel.

Our main result is the same as that given recently in independent work by Holevo [6]. Holevo's proof, like ours, follows the general strategy of [5], though there are substantial differences of detail.

II. SETTING IT UP

We will assume that we have an alphabet of mixed states W_x, each of which has an *a priori* probability p_x. The average density matrix is $W=\Sigma_x p_x W_x$. We wish to show that, if we use long strings of these letters (suitably pruning the set of codewords to improve distinguishability) and an appropriate decoding observable, we can send reliably an amount of information up to

$$\chi = H(W) - \sum_x p_x H(W_x) \tag{2}$$

per letter.

We will be considering strings of L letters. In what follows we will assume that the index a refers to a whole string of letters: $a = x_1 \cdots x_L$. $P_a = p_{x_1} \cdots p_{x_L}$ is the *a priori* probability of the sequence a and $\rho_a = W_{x_1} \otimes \cdots \otimes W_{x_L}$ is the state associated with the string. The average state is

$$\rho = \sum_a P_a \rho_a = \underbrace{W \otimes \cdots \otimes W}_{L \text{ times}}. \tag{3}$$

Consider the state ρ_a. This has a complete orthogonal set of eigenstates, which we will denote $|s_{ak}\rangle$ (where k ranges over the dimension of the space), and a corresponding set of eigenvalues $p_{k|a}$. As the notation suggests, we may think of the $p_{k|a}$'s as "conditional probabilities" for k given a, and this motivates us to form the "joint probability" distribution $P_{ak} = P_a p_{k|a}$. Of course, $\rho = \Sigma_{ak} P_{ak} |s_{ak}\rangle\langle s_{ak}|$. It will be convenient to refer to the index k identifying the string eigenstate as the *syndrome* of the codeword ρ_a.

When we construct our decoding observable, we will be trying not only to distinguish the string a, but also (a seemingly harder task) to determine the syndrome k as well. An error will occur if either the codeword or the syndrome is incorrectly identified. For a given codeword ρ_a, the various $|s_{ak}\rangle$'s are orthogonal and hence perfectly distinguishable from one another, so this will not really be more difficult than identifying the codeword only.

III. TYPICALITY

Let $\epsilon, \delta > 0$. Then we can find a length L large enough to enforce the following typicality conditions on strings of length L.

(i) There exists a typical subspace [8,9] for the states. That is, there is a subspace Λ spanned by eigenstates of ρ such that, if Π is the projection onto Λ, $\mathrm{Tr}\rho\Pi > 1-\epsilon$. Further, if we denote by $|\lambda_n\rangle$ the eigenstate of ρ with eigenvalue λ_n,

$$2^{-L[H(W)+\delta]} < \lambda_n < 2^{-L[H(W)-\delta]} \tag{4}$$

for all $|\lambda_n\rangle \in \Lambda$.

One key property of the typical subspace is that

$$\mathrm{Tr}\rho^2\Pi < 2^{-L[H(W)-2\delta]}. \tag{5}$$

This property was used by Hausladen *et al.* [5] to bound the probability of error, and it will play that role again.

(ii) There exist a typical set of strings (relative to the distribution P_a) and a typical set of string-syndrome pairs (relative to the joint distribution P_{ak}). Let $H(A)$ be the Shannon entropy associated with the string distribution P_a and let $H(A,K)$ be the Shannon entropy associated with the joint distribution P_{ak}. Notice that $H(A) = LH(X)$, where $H(X)$ is the Shannon entropy of the letter distribution. Also,

$$\begin{aligned} H(A,K) &= H(A) + \sum_a P_a H(\rho_a) \\ &= L\left(H(X) + \sum_x p_x H(W_x) \right), \end{aligned} \tag{6}$$

where the x sum is over the letters. Typicality means the following.

(a) For a typical string a,

$$2^{-L[H(X)+\delta]} < P_a < 2^{-L[H(X)-\delta]}. \tag{7}$$

Furthermore, the sum of the P_a's for the typical strings is greater than $1-\epsilon$.

(b) For a typical string-syndrome pair ak,

$$\begin{aligned} 2\ \exp\left[-L\left(H(X) + \sum_x p_x H(W_x) + \delta \right) \right] &< P_{ak} \\ < 2\ \exp\left[-L\left(H(X) + \sum_x p_x H(W_x) - \delta \right) \right], \end{aligned} \tag{8}$$

where $2\ \exp[x]$ means 2^x. Furthermore, the sum of P_{ak} over the typical string-syndrome pairs is also greater than $1-\epsilon$.

For each string a, we define a set of *relatively typical* syndromes as follows: k is relatively typical to a if a is a typical string and ak is a typical string-syndrome pair. (Note that atypical strings have no relatively typical syndromes.) If k is relatively typical to a, then

$$\begin{aligned} 2\ \exp\left[-L\left(\sum_x p_x H(W_x) + 2\delta \right) \right] &< p_{k|a} \\ < 2\ \exp\left[-L\left(\sum_x p_x H(W_x) - 2\delta \right) \right] \end{aligned} \tag{9}$$

since $p_{k|a} = P_{ak}/P_a$. We can take advantage of the definition of χ above to write this as

$$2^{-L[H(W)-\chi+2\delta]} < p_{k|a} < 2^{-L[H(W)-\chi-2\delta]}. \tag{10}$$

We adopt the following notations for sums: $\Sigma_{k|a}$ means sum over k for a given value of a, $\Sigma'_{k|a}$ means sum restricted to relatively typical k's only (note that this sum may have no terms), and Σ'_{ak} means $\Sigma_a \Sigma'_{k/a}$. If we restrict sums to relatively typical syndromes only, we do not lose much weight in the ensemble. That is, consider the pairs ak in which k is relatively typical. This excludes all atypical a's (a set of total probability less than ϵ) and all atypical pairs ak (also of probability less than ϵ). It follows that

$$\sum_{ak}{}' P_{ak} = \sum_a P_a \sum_{k|a}{}' p_{k|a} > 1 - 2\epsilon. \tag{11}$$

The total ensemble $\rho = \Sigma_{ak} P_{ak} |s_{ak}\rangle\langle s_{ak}|$. If we restrict the ensemble to string-syndrome pairs in which the syndrome is relatively typical, then we get a subnormalized density operator $\tilde{\rho}$ for which

$$\mathrm{Tr}\tilde{\rho} = \mathrm{Tr}\left(\sum_{ak} P_{ak} |s_{ak}\rangle\langle s_{ak}| \right) > 1 - 2\epsilon. \tag{12}$$

We also note that $\tilde{\rho} \leq \rho$ under the usual partial ordering of positive operators. (That is, $\langle\psi|\tilde{\rho}|\psi\rangle \leq \langle\psi|\rho|\psi\rangle$ for all $|\psi\rangle$.)

IV. CODING AND DECODING

Now we discuss our code and our decoding procedure. The code will consist of N codewords (each a string of length L), which we will use with equal frequency. Codewords in our code will be indexed by a greek index such as α. Thus the latin characters $a, b, \ldots$ index the whole set of strings, while the greek characters $\alpha, \beta, \ldots$ index the codewords in our code. Greek indices thus take on N possible values.

The decoding procedure will be a variation of the ‘‘pretty good measurement’’ used in [5]. We will attempt to identify not only the codeword but also the syndrome. Our decoding observable will be a ‘‘positive operator measurement’’ (POM), described by a set of positive operators summing to unity. For each codeword-syndrome pair αk we will have a (possibly subnormalized) vector $|\tilde{\mu}_{\alpha k}\rangle$ such that $|\tilde{\mu}_{\alpha k}\rangle\langle\tilde{\mu}_{\alpha k}|$ is an element of our decoding POM. The probability of error is thus

$$\begin{aligned} P_E &= 1-\frac{1}{N}\sum_{\alpha k} p_{k|\alpha}|\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle|^2 \\ &= \frac{1}{N}\sum_{\alpha}\sum_{k\alpha} p_{k|\alpha}(1-|\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle|^2) \\ &\leqslant 2\left(1-\frac{1}{N}\sum_{\alpha k} p_{k|\alpha}\left|\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle\right|\right). \end{aligned} \tag{13}$$

We next describe how to specify the decoding observable. If k is not relatively typical to α, we let $|\tilde{\mu}_{\alpha k}\rangle=0$. For the rest, we construct the operator

$$\Upsilon=\sideset{}{'}\sum_{\alpha k}\Pi|s_{\alpha k}\rangle\langle s_{\alpha k}|\Pi, \tag{14}$$

where Π is the projection onto the typical subspace Λ for the *a priori* ensemble, as described above. We define

$$|\tilde{\mu}_{\alpha k}\rangle=\Upsilon^{-1/2}|s_{\alpha k}\rangle \tag{15}$$

for k relatively typical to α. (Since Υ is not generally fully invertible, this $\Upsilon^{-1/2}$ is the pseudoinverse of $\Upsilon^{1/2}$, supported only on the support of Υ.) It follows that

$$\sum_{\alpha,k}|\tilde{\mu}_{\alpha k}\rangle\langle\tilde{\mu}_{\alpha k}|=\sideset{}{'}\sum_{\alpha k}|\tilde{\mu}_{\alpha k}\rangle\langle\tilde{\mu}_{\alpha k}|=1 \tag{16}$$

on the range space of Υ (which is a subspace of Λ). We can add an element of the POM (labeled ‘‘error’’) on the orthogonal space, if necessary, to give overall normalization.

Since Υ (and thus $\Upsilon^{-1/2}$) is positive, the inner product $\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle$ is real and non-negative. We do not need the modulus signs in our bound for the probability of error. Furthermore, our construction of the $|\tilde{\mu}_{\alpha k}\rangle$'s means that the only contributions come from those terms in which k is relatively typical to α. Thus we can write

$$P_E\leqslant 2\left(1-\frac{1}{N}\sideset{}{'}\sum_{\alpha k} p_{k|\alpha}\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle\right). \tag{17}$$

As was mentioned in [5], this definition has some nice properties connected with a matrix of inner products of $|s_{\alpha k}\rangle$. For k relatively typical to α and l relatively typical to β, we define

$$\mathcal{S}_{\alpha k,\beta l}=\langle s_{\alpha k}|\Pi|s_{\beta l}\rangle. \tag{18}$$

The matrix $\mathcal{S}$ is a positive square matrix. It turns out that

$$\langle\tilde{\mu}_{\alpha k}|s_{\alpha k}\rangle=(\sqrt{\mathcal{S}})_{\alpha k,\alpha k}. \tag{19}$$

(This could be used as an implicit definition of the $|\tilde{\mu}_{\alpha k}\rangle$'s.) We will employ the same inequality that was used in [5]: for $x\geqslant 0$, $\sqrt{x}\geqslant\frac{3}{2}x-\frac{1}{2}x^2$. This means that

$$(\sqrt{\mathcal{S}})_{\alpha k,\alpha k}\geqslant\frac{3}{2}\mathcal{S}_{\alpha k,\alpha k}-\frac{1}{2}\sideset{}{'}\sum_{\beta,l}\mathcal{S}_{\alpha k,\beta l}\mathcal{S}_{\beta l,\alpha k}. \tag{20}$$

This gives us a bound for the probability of error

$$\begin{aligned} P_E \le 2 &- \underbrace{\frac{3}{N}\sum_{\alpha}\sideset{}{'}\sum_{k|\alpha} p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\alpha k}\rangle}_{\#1} \\ &+ \underbrace{\frac{1}{N}\sum_{\alpha\beta}\sideset{}{'}\sum_{k|\alpha}\sideset{}{'}\sum_{l|\beta} p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\beta l}\rangle\langle s_{\beta l}|\Pi|s_{\alpha k}\rangle}_{\#\ 2} \end{aligned} \tag{21}$$

We will deal with the terms labeled # 1 and # 2 separately.

V. RANDOM CODES

Now we will average the probability of error P_E over random codes. These codes are constructed by choosing the N codewords independently according to the *a priori* string distribution P_a. This will have the effect of turning averages over the codewords in the code into averages over the *a priori* string ensemble.

Denote the random code average by $\langle\ \rangle_c$. Consider term #1 above,

$$\begin{aligned} \langle\#1\rangle_c &= \left\langle\frac{3}{N}\sum_{\alpha}\sideset{}{'}\sum_{k|\alpha} p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\alpha k}\rangle\right\rangle_c \\ &= 3\left(\sum_{a} P_a\sideset{}{'}\sum_{k|a} p_{k|a}\mathrm{Tr}\Pi|s_{ak}\rangle\langle s_{ak}|\Pi\right) \\ &= 3(\mathrm{Tr}\Pi\tilde{\rho}\Pi). \end{aligned} \tag{22}$$

(Notice that the average over random codes transformed the sum over the codewords Σ_α into N times the average over the string ensemble described by P_a since each codeword is chosen independently according to P_a.) Now let $\Delta=\rho-\tilde{\rho}$, a positive operator since $\tilde{\rho}\leqslant\rho$. Then

$$\begin{aligned} \langle\#1\rangle_c &= 3(\mathrm{Tr}\Pi\rho\Pi-\mathrm{Tr}\Pi\Delta\Pi)>3[(1-\epsilon)-\mathrm{Tr}\Delta] \\ &> 3(1-3\epsilon) \end{aligned} \tag{23}$$

since $\mathrm{Tr}\Delta<2\epsilon$.

Next, examine term #2. The double sum over α and β may be split into two parts: a part in which $\alpha=\beta$ and a part in which $\alpha\neq\beta$. The advantage in this is that, if $\alpha\neq\beta$, the codewords are chosen independently in a random code:

$$\#2=\underbrace{\frac{1}{N}\sum_{\alpha}\overset{\prime}{\sum_{k|\alpha}}\overset{\prime}{\sum_{l|\alpha}}p_{k|\alpha}\langle s_{\alpha k}\,|\,\Pi\,|\,s_{\alpha l}\rangle\,\langle s_{\alpha l}\,|\,\Pi\,|\,s_{\alpha k}\rangle}_{\#\ 2\mathrm{a}}$$
$$+\underbrace{\frac{1}{N}\sum_{\alpha,\beta\neq\alpha}\overset{\prime}{\sum_{k|\alpha}}\overset{\prime}{\sum_{l|\beta}}p_{k|\alpha}\langle s_{\alpha k}\,|\,\Pi\,|\,s_{\beta l}\rangle\,\langle s_{\beta l}\,|\,\Pi\,|\,s_{\alpha k}\rangle\,.}_{\#\ 2\mathrm{b}}$$
$$(24)$$

We consider term $\#2a$:

$$\#2a=\frac{1}{N}\sum_{\alpha}\ \overset{\prime}{\sum_{k|\alpha}}\ \overset{\prime}{\sum_{l|\alpha}}\ p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\alpha l}\rangle\langle s_{\alpha l}|\Pi|s_{\alpha k}\rangle$$
$$\leqslant\frac{1}{N}\overset{\prime}{\sum_{\alpha}}\ \overset{\prime}{\sum_{k|\alpha}}\ \sum_{l|a}\ p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\alpha l}\rangle\langle s_{\alpha l}|\Pi|s_{\alpha k}\rangle$$
$$=\frac{1}{N}\sum_{\alpha}\ \overset{\prime}{\sum_{k|\alpha}}\ p_{k|\alpha}\left\langle s_{\alpha k}\middle|\Pi\left(\sum_{l|a}\ |s_{\alpha l}\rangle\langle s_{\alpha l}|\right)\Pi\middle|s_{\alpha k}\right\rangle$$
$$=\frac{1}{N}\sum_{\alpha}\ \overset{\prime}{\sum_{k|\alpha}}\ p_{k|\alpha}\langle s_{\alpha k}|\Pi|s_{\alpha k}\rangle \qquad (25)$$

since for any α, the $|s_{\alpha l}\rangle$ form a complete set. But $\langle s_{\alpha k}|\Pi|s_{\alpha k}\rangle\leqslant 1$, so

$$\#2a\leqslant\frac{1}{N}\sum_{\alpha}\ \overset{\prime}{\sum_{k|\alpha}}\ p_{k|\alpha}\leqslant\frac{1}{N}\sum_{\alpha}\ \sum_{k\alpha}\ p_{k|\alpha}=1. \qquad (26)$$

Therefore, of course, $\langle\#2a\rangle_c\leqslant 1$.

Now we consider the much more interesting term #2b:

$$\#2b=\frac{1}{N}\sum_{\alpha,\beta\neq\alpha}\ \sum_{k|\alpha}\ \sum_{l|\beta}\ p_{k|\alpha}\mathrm{Tr}\Pi|s_{\alpha k}\langle s_{\alpha k}|\Pi|s_{\beta l}\rangle\langle s_{\beta l}|\Pi. \qquad (27)$$

The only terms that appear in this sum are terms in which l is typical relative to β. But for such codeword-syndrome pairs, we have a uniform lower bound on $p_{l|\beta}$, which allows us to say that, for all l and β that appear in our sum,

$$1<p_{l|\beta}2^{L[H(W)-\chi+2\delta]}. \qquad (28)$$

Therefore,

$$\#2b\leqslant 2^{L[H(W)-\chi+2\delta]}\frac{1}{N}\sum_{\alpha,\beta\neq\alpha}\ \overset{\prime}{\sum_{k|\alpha}}\ \overset{\prime}{\sum_{l|\beta}}$$
$$\times p_{k|\alpha}p_{l|\beta}\mathrm{Tr}\Pi|s_{\alpha k}\rangle\langle s_{\alpha k}|\Pi|s_{\beta l}\rangle\langle s_{\beta l}|\Pi. \qquad (29)$$

Taking the average of $\#2b$ over random codes,

$$\langle\#2b\rangle_c=2^{L[H(W)-\chi+2\delta]}\frac{N(N-1)}{N}$$
$$\times\left(\sum_{a}\ P_a\sum_{b}\ P_b\overset{\prime}{\sum_{k|a}}\ p_{k|a}\overset{\prime}{\sum_{l|b}}\ p_{l|b}\right.$$
$$\left.\times\mathrm{Tr}\Pi|s_{ak}\rangle\langle s_{ak}|\Pi|s_{bl}\rangle\langle s_{bl}|\Pi\right)$$
$$\leqslant N2^{L[H(W)-\chi+2\delta]}\mathrm{Tr}\Pi\tilde{\rho}\Pi\tilde{\rho}\Pi. \qquad (30)$$

(Notice again that each term in the sums over the codewords has been replaced by the appropriate string-ensemble average.) We note that if A, B, and C are positive operators with $B\leqslant A$, $\mathrm{Tr}BC\leqslant\mathrm{Tr}AC$. Thus

$$\mathrm{Tr}\Pi\tilde{\rho}\Pi\tilde{\rho}\Pi=\mathrm{Tr}\Pi\tilde{\rho}\Pi\tilde{\rho}\leqslant\mathrm{Tr}\Pi\tilde{\rho}\Pi\rho=\mathrm{Tr}\tilde{\rho}\Pi\rho\Pi\leqslant\mathrm{Tr}\rho\Pi\rho\Pi$$
$$=\mathrm{Tr}\rho^2\Pi, \qquad (31)$$

where the last line uses the fact that ρ and Π commute:

$$\langle\#2b\rangle_c\leqslant N2^{L[H(W)-\chi+2\delta]}\mathrm{Tr}\rho^2\Pi$$
$$<N2^{L[H(W)-\chi+2\delta]}2^{-L[H(W)-2\delta]}=N2^{-L(\chi-4\delta)}.$$
$$(32)$$

Combining these results, we can find an upper bound for the probability of error averaged over all random codes:

$$\langle P_E\rangle_c\leqslant 2-\langle\#1\rangle_c+\langle\#2a\rangle_c+\langle\#2b\rangle_c$$
$$<2-3(1-3\epsilon)+1+N2^{-L(\chi-4\delta)}=9\epsilon+N2^{-L(\chi-4\delta)}.$$
$$(33)$$

For L sufficiently large, we can choose N nearly as big as $2^{L\chi}$ and still have the probability of error small.

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

$$P_E\leqslant 9\epsilon+N2^{-L(\chi-4\delta)}. \qquad (34)$$

If L is very large, Alice can use up to $N=2^{L(\chi-5\delta)}$ codewords and still have $P_E\leqslant 10\epsilon$. In this case, Alice encodes $\chi-5\delta$ bits per letter. This proves our main theorem.

We have shown the existence of a satisfactory code without actually constructing it. Consequently, we do not know much about the structure of the code. In particular, we have not guaranteed in our proof that the letter states occur in the codewords with frequencies that closely match their *a priori* probabilities p_x. (This is something that we might wish to require since the distribution p_x might be chosen to optimize some resource, such as the energy required per letter.) It turns out, however, that we can satisfy such a requirement. Since we generate the codewords in our ensemble of codes by using the *a priori* probabilities, the law of large numbers implies that the letter frequencies will match the *a priori* distribution within any specified tolerance for a set of ‘‘typical codes.’’ The set of typical codes includes almost the entire weight of the code ensemble and thus many of the particular codes with low probability of error. See [5] for the details of this argument applied to the pure state case.

VI. FIXED-ALPHABET CAPACITY

We have shown that it is possible to send information at any rate up to χ bits per letter with arbitrarily low probability of error. The capacity of a channel is defined as the maximum information per letter that may be sent through the channel with P_E arbitrarily small. Thus χ provides a lower bound to the capacity of the quantum channel.

Classical information theory together with Kholevo's theorem also allows us to use χ to establish an *upper bound* for the capacity of the channel. Suppose X represents Alice's input and Y represents Bob's decoding measurement outcome. Then the Fano inequality [7] states that

$$\begin{aligned}-P_E\log_2 P_E-(1-P_E)\log_2(1-P_E)+P_E\log_2(N_X-1)\\ \geqslant H(X|Y), \qquad (35)\end{aligned}$$

where P_E is the probability of error and N_X is the number of possible values of X. $H(X|Y)$ is the conditional Shannon entropy of X given Y, that is, the entropy of the conditional distribution $p(x|y)$, averaged over the various values of y [13]. It is related to the mutual information $H(X:Y)$ by

$$H(X|Y)=H(X)-H(X:Y). \qquad (36)$$

In the channel, Alice uses some signal states ρ_a with probabilities P_a. Kholevo's theorem places an upper bound on the mutual information $H(X:Y)$:

$$H(X:Y)\leqslant H(\rho)-\sum_a P_a H(\rho_a).$$

(Note that if the channel used by Alice and Bob consists of L letters used independently, then the Kholevo bound is just $L\chi$, where χ is the Kholevo bound for a single letter.) If the Alice's input X has an entropy $H(X)$ that exceeds $H(\rho)-\Sigma_a P_a H(\rho_a)$, then $H(X|Y)>0$ and it will not be possible to make the probability of error P_E arbitrarily small.

Suppose we fix an alphabet $\Gamma=\{W_x\}$ of letter states W_x and require that Alice use codewords a that are length-L strings of these letter states: $a=x_1\cdots x_L$. Then the probability distribution P_a yields marginal probability distributions $p(x_1),\ldots,p(x_L)$ and average density operators $W_1,\ldots,W_L$ for the L different letters. It follows that

$$\begin{aligned}&H(\rho)-\sum_a P_a H(\rho_a)\\ &=H(\rho)-\sum_a P_a(H(W_{x_1})+\cdots+H(W_{x_L}))\\ &=H(\rho)-\left(\sum_{x_1} p(x_1)H(W_{x_1})+\cdots\right.\\ &\quad\left.+\sum_{x_L} p(x_L)H(W_{x_L})\right)\\ &\leqslant\left(H(W_1)-\sum_{x_1} p(x_1)H(W_{x_1})\right)+\cdots\\ &\quad+\left(H(W_L)-\sum_{x_L} p(x_L)H(W_{x_L})\right), \qquad (37)\end{aligned}$$

where we have used the subadditivity of the entropy $H(\rho)$. We might write this as

$$\chi^{(L)}\leqslant\chi_1+\cdots+\chi_L, \qquad (38)$$

where $\chi^{(L)}$ represents the Kholevo bound for the ensemble of codewords of length L and $\chi_1,\ldots,\chi_L$ represent Kholevo bounds for the individual letter ensembles.

We define the *fixed-alphabet capacity* C_Γ to be

$$C_\Gamma=\sup_{p(x)}\chi, \qquad (39)$$

where $p(x)$ is the probability distribution over the letter states in Γ and χ is the single-letter Kholevo bound. This quantity represents the maximum information rate per letter that Alice can send to Bob with arbitrarily low probability of error.

This claim follows directly from our results so far. Suppose Alice uses codewords of length L. Then $\chi^{(L)}\leqslant LC_\Gamma$; by the above argument, if Alice attempts to send more than LC_Γ bits using these codewords then the probability of error will not be arbitrarily small. Conversely, we can choose the letter probabilities so that χ is as close as required to C_Γ, and we have previously shown that a suitable choice of code and decoding observable can convey up to χ bits per letter with arbitrarily low P_E. Thus the capacity C_Γ cannot be exceeded, but can be approached arbitrarily closely.

VII. NOISY CHANNELS

The mixed states W_x used in our alphabet are the states available to Bob for decoding. They may in fact not be the original states of the channel Q chosen by Alice. In the interval between Alice's encoding and Bob's decoding, the system Q may have undergone unitary internal evolution (which Bob can correct by a suitable choice of "rotated" decoding observable) and interaction with the external environment (which Bob cannot in general correct).

The most general description of the evolution of a quantum system Q interacting with an environment is provided by a trace-preserving completely positive linear map on the set of density operators of Q [11]. Such a map is described by a superoperator $\mathcal{E}$:

$$\rho\rightarrow\rho'=\mathcal{E}(\rho), \qquad (40)$$

where ρ is the initial state of the system and ρ' is the final state. The superoperator $\mathcal{E}$ acts linearly, so that a convex combination of input states yields a convex combination of output states. This description clearly includes unitary evolution of Q as a special case, but it also can account for interaction with the environment.

A noisy quantum channel is defined by a superoperator $\mathcal{E}$ that describes the evolution of each letter as it is transmitted from Alice to Bob. We assume that the channel is memoryless, i.e., that the evolution of each letter is independent.

This means, among other things, that a product state of several input letters will evolve into a product state output.

Alice's basic problem is to use input states w_x so that the output states $W_x = \mathcal{E}(w_x)$ can be distinguished by Bob. If Alice has a fixed alphabet $\{w_x\}$ of input states, then the maximum achievable information rate per letter is still given by our fixed-alphabet capacity C_Γ, where Γ is the alphabet of *output* states.

Now suppose that Alice is allowed to choose her input states in order to maximize the information conveyed to Bob over the noisy quantum channel, subject to the constraint that Alice must transmit codewords that are represented by product states of the letters. This *almost* reduces to the fixed-alphabet problem, where the fixed alphabet Γ now includes all of the possible output states of the channel. The maximum over probability distributions is now a maximum over all input ensembles of states chosen by Alice.

We say that this problem *almost* reduces to the fixed alphabet problem in that the argument that χ is an upper bound of the capacity must be modified in this case. Recall from Sec. VI that we applied the classical Fano inequality to show that if Alice attempts to send information at a rate exceeding χ, then the probability of error cannot be made arbitrarily small. If we attempt to use the same argument in the present case, then the Fano inequality does not help us for at least two reasons. First, the number of possible input states N_X is unbounded. Second, we do not have a characterization of $H(X|Y)$ that allows us to compare it with N_X. Thus we will modify the Fano inequality to understand the behavior of the probability of error in the present case.

We first note that the probability of ''getting it right''

$$1 - P_E = \frac{1}{N} \sum_{\alpha k} p_{k|\alpha} |\langle \tilde{\mu}_{\alpha k} | s_{\alpha k} \rangle|^2 \tag{41}$$

is linear in the elements of the POM. Thus the probability of error P_E is a convex function on the elements of the POM. We may modify the proof of a result of Davies (Theorem 3 of [14]) to show that the convex function P_E is minimized by a POM having no more than d^2 elements, where d is the dimension of the support of the POM. Thus the probability of error is minimized by a decision scheme in which at most d^2 of the inputs are identified by the decision scheme. Let us denote the output of such a scheme by $Y_{\min}$. Fano's inequality gives us that

$$-P_E \log_2 P_E - (1-P_E)\log_2(1-P_E) + P_E \log_2(d^2-1) \geqslant H(X|Y_{\min}). \tag{42}$$

Note that

$$H(X|Y_{\min}) = H(X) - H(X{:}Y_{\min}\mathrm{min}) \tag{43}$$

$$\geqslant H(X) - \chi, \tag{44}$$

so that we conclude

$$-P_E \log_2 P_E - (1-P_E)\log_2(1-P_E) + P_E \log_2(d^2-1) \geqslant H(X) - \chi. \tag{45}$$

Note that this is a relation between the minimum probability of error and a quantity $[H(X)-\chi]$ that does not depend on the particular decision scheme. We see that if Alice attempts to send information at a rate $H(X)$ in excess of χ, then the probability of error cannot be made arbitrarily small.

We now turn to a demonstration that this rate can be achieved. Alice wishes to choose a set of input states w_x (together with input probabilities p_x) so that χ is maximized for the output states W_x. We next show that Alice can do no better than choose the input states w_x to be pure. Let a set of (possibly mixed) input states w_x be given along with their *a priori* probabilities and let

$$W = \sum_x p_x W_x = \sum_x p_x \mathcal{E}(w_x) \tag{46}$$

be the average output state. Then

$$\chi = H(W) - \sum_x p_x H(\mathcal{E}(w_x)). \tag{47}$$

Construct a new set of pure state inputs by resolving each mixed state input into a convex combination of pure states:

$$w_x = \lambda_{xk} |\psi_{xk}\rangle\langle\psi_{xk}|. \tag{48}$$

We will use the state $|\psi_{xk}\rangle$ with probability $p_{xk} = p_x \lambda_{xk}$. By linearity,

$$W_x = \mathcal{E}(w_x) = \sum_k \lambda_{xk} \mathcal{E}(|\psi_{xk}\rangle\langle\psi_{xk}|), \tag{49}$$

so that the average output state is still W, as before. By the convexity of the von Neumann entropy,

$$H(W_x) \geqslant \sum_k \lambda_{xk} H(\mathcal{E}(|\psi_{xk}\rangle\langle\psi_{xk}|)). \tag{50}$$

It follows that

$$\chi' = H(W) - \sum_{xk} p_{xk} H(\mathcal{E}(|\psi_{xk}\rangle\langle\psi_{xk}|)) \geqslant H(W) - \sum_x p_x H(\mathcal{E}(w_x)) = \chi. \tag{51}$$

In other words, for any ensemble of mixed input states, we can find an ensemble of pure input states whose output states have a χ at least as great. The optimal inputs for the noisy quantum channel are pure states.

To sum up, if Alice is required to use product states to represent her codewords, then the capacity $C^{(1)}$ of the noisy quantum channel is

$$C^{(1)} = \max \chi, \tag{52}$$

where χ is the Kholevo bound for the output states of the channel and the maximum is taken over all ensembles of pure state inputs. Alice can reliably transmit information to

Bob at any rate below $C^{(1)}$. We will refer to $C^{(1)}$ as the product state capacity. The superscript (1) reminds us that Alice is required to use the multiple available copies of the channel *one at a time*, coding her messages into product states.

The product state capacity $C^{(1)}$ is a function only of the superoperator $\mathcal{E}$ describing the dynamical evolution of a single channel. To emphasize this, we will calculate $C^{(1)}$ in the simple case of a one-quantum-bit (one-qubit) depolarizing channel. A two-level system, or qubit, is sent through the channel. With probability P, the state of the qubit is left intact; with probability $1-P$, the state is completely randomized, so that the output state of the qubit is a completely mixed density operator. For any pure state input $w_x=|\psi_x\rangle\langle\psi_x|$, the output state is

$$W_x=\mathcal{E}(w_x)=P|\psi_x\rangle\langle\psi_x|+\frac{1-P}{2}I, \tag{53}$$

where I is the identity operator. Any such state has eigenvalues $\frac{1}{2}(1+P)$ and $\frac{1}{2}(1-P)$ and thus an entropy

$$\begin{aligned}H(W_x)&=-\frac{1+P}{2}\log_2\frac{1+P}{2}-\frac{1-P}{2}\log_2\frac{1-P}{2}\\&=1-\frac{1}{2}[(1+P)\log_2(1+P)+(1-P)\log_2(1-P)].\end{aligned} \tag{54}$$

To calculate the capacity, we maximize the output χ over all ensembles of pure state inputs. But the entropy of each output state will be the same, so we only need to maximize the entropy of the average output state W. This is easily seen to be 1, so that

$$C^{(1)}=\frac{1}{2}[(1+P)\log_2(1+P)+(1-P)\log_2(1-P)]. \tag{55}$$

If $P=0$, then the product state capacity is (reassuringly) zero; but for any $P>0$, the product state capacity $C^{(1)}>0$, with $C^{(1)}=1$ bit for $P=1$.

However, Alice can do more than we have so far allowed her to do. It might conceivably be to her advantage to use *entangled* states to represent her codewords. The output states will in general be entangled states. (This will present no additional difficulties for Bob; even to distinguish product states, we have allowed Bob to use a collective decoding observable for strings of L letters.)

In this case, it is no longer true that the Kholevo bound $\chi^{(L)}$ for the output codewords satisfies

$$\chi^{(L)}\leqslant\chi_1+\cdots+\chi_L, \tag{56}$$

where the χ_k denote the Kholevo bounds for the individual letters. That is, χ is not necessarily subadditive for systems that may be entangled.

Suppose that Alice is permitted to prepare entangled states of L copies of the channel. Then we can treat these L copies as a single ''extended'' channel, which Alice can prepare in any state. Our main theorem applied to this extended channel means that for any $\chi^{(L)}$ of the output states, Alice can reliably send up to $\chi^{(L)}/L$ bits of information per letter to Bob. Thus we define

$$C^{(L)}=\frac{1}{L}\max\chi^{(L)}, \tag{57}$$

where the maximum is taken over all input ensembles, including entangled states, for the L elementary channels. (By our previous arguments, it suffices to consider only ensembles of pure input states.) $C^{(L)}$ is the capacity if Alice is allowed to use the channels in entangled blocks of length L. Since product states are allowed, it is clear that $C^{(L)}\geqslant C^{(1)}$. The asymptotic capacity will be

$$C=\lim_{L\to\infty}C^{(L)}. \tag{58}$$

This will be the ultimate information capacity of the noisy quantum channel. (Similar considerations are discussed in [10].)

Like $C^{(1)}$, C will be a function only of the dynamical superoperator $\mathcal{E}$. No examples are known where $C>C^{(1)}$ (though the example in [12] is suggestive). Thus it is not known whether or not $C=C^{(1)}$.

ACKNOWLEDGMENTS

The authors wish to thank L. Levitin, R. Jozsa, and W. Wootters for many helpful discussions and suggestions about this and related work. The authors also wish to thank the Institute for Scientific Interchange in Turin, Italy for the opportunities afforded by a workshop on quantum computation. M.D.W. was supported by the Robert C. Good Foundation at Denison University.

[1] 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.

[2] L. B. Levitin, *Information, Complexity, and Control in Quantum Physics*, edited by A. Blaquière, S. Diner, and G. Lochak (Springer, Vienna, 1987), pp. 111–115.

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

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

[5] P. Hausladen, R. Josza, B. Schumacher, M. Westmoreland, and W. K. Wootters, Phys. Rev. A **54**, 1869 (1996).

[6] A. S. Kholevo, IEEE Trans. Inf. Theory (to be published).

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

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

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

[10] A. S. Kholevo, Probl. Peredachi Inf. **15**, 3 (1979).

[11] K. Hellwig and K. Kraus, Commun. Math. Phys. **16**, 142 (1970); M.-D. Choi, Linear Algebr. Appl. **10**, 285 (1975); K. Kraus, *States Effects and Operations: Fundamental Notions of Quantum Theory* (Springer-Verlag, Berlin, 1983).

[12] C. H. Bennett, C. A. Fuchs, and J. Smolin (unpublished).

[13] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).

[14] E. B. Davies, IEEE Trans. Inf. Theory **IT-24**, 596 (1978).