Capacities of Quantum Channels and How to Find Them

Peter W. Shor
AT&T Labs – Research
Florham Park, NJ 07922

Abstract: We survey what is known about the information transmitting capacities of quantum channels, and give a proposal for how to calculate some of these capacities using linear programming.

1 Introduction

In this paper, we discuss the capacity of quantum channels. Information theory says that the capacity of a classical channel is essentially unique, and is representable as a single numerical quantity, which gives the amount of information that can be transmitted asymptotically per channel use [46, 15]. Quantum channels, unlike classical channels, do not have a single numerical quantity which can be defined as their capacity for transmitting information. Rather, quantum channels appear to have at least four different natural definitions of capacity, depending on the auxiliary resources allowed, the class of protocols allowed, and whether the information to be transmitted is classical or quantum.

In this paper, we first introduce the background necessary for understanding the capacity of quantum channels, and then define several capacities of these channels. For two of these channel capacities, we sketch possible techniques for computing them which we believe will be more efficient than techniques currently used. These capacities are both reducible to optimization problems over matrices. We believe that a combination of linear programming techniques, including column generation, and non-linear optimization will provide a more efficient method for calculating these capacities. Unfortunately, at the time of writing this paper, I have not yet tested these techniques experimentally. Since I cannot prove that these techniques are efficient, the proof of this pudding must be in the computing, and is thus not yet demonstrated. We hope to test these techniques in the near future.

To date, the means used for numerical computations of quantum channel capacities have been fairly straightforward, often using gradient descent techniques [40]. More research has been done on the calculation of the entanglement of formation [51, 4], a related problem [36]. None of these programs have used combinatorial optimization techniques. For one of the capacities discussed in this paper—the entanglement-assisted capacity—this technique may be fairly efficient, as this capacity has a single local optimum which is also a global optimum. For two other capacities discussed in this paper—the $C_{1,1}$ and $C_{1,\infty}$ capacities—I propose techniques involving linear programming that could be used for the capacity computation, and which I suspect are much more efficient than straightforward optimization. For another capacity—the one-way quantum capacity—there are multiple local maxima in the optimization problem, and we need to determine the global maximum. In this case, unfortunately, although hill climbing does not seem like it would be efficient, I do not have any alternative techniques to suggest.

This paper originates in my research investigating the capacities of a quantum channel [48]. In order to show that a certain channel capacity (which I do not deal with in this paper; it is less natural than the capacities covered here) lies strictly between two other channel capacities, I needed to calculate some of these capacities. Specifically, I needed to calculate what I call the $C_{1,1}$ capacity of a fairly simple quantum channel. I realized that this was a problem which could be solved numerically using linear programming, and I used this technique to obtain a picture of the $C_{1,1}$ capacity landscape which was satisfactory for my application. During this computation, it became clear that a better way to solve this problem would be to use column generation techniques to make the linear program more efficient, and that these would furthermore also be useful for computing other capacities of quantum channels. I have not yet had time to experimentally test these new techniques (rather, my program started with enough columns to ensure obtaining a close approximation of the capacity; this would be an enormous waste of resources for larger problems, but for my purposes it was quite adequate). This paper will explain the column generation technique. I will try to make it comprehensible both to researchers with background in mathematical programming and to researchers with background in quantum information theory. Those wishing more background information on linear programming, on quantum computing and information, or on classical information theory can find them in textbooks such as [14, 38, 15] More specifically, I will give proposals for how to compute two capacities for carrying classical information over a quantum channel: namely, the $C_{1,1}$ capacity and the $C_{1,\infty}$ capacity. These techniques should also work for computing a formula that I conjecture gives the classical entanglement-assisted capacity with limited entanglement; this extrapolates between the $C_{1,\infty}$ capacity and the entanglement-assisted capacity. The description of quantum information theory and capacities contained in here is largely taken from the paper [47].

2 Quantum Information Theory

The discipline of information theory was founded by Claude Shannon in a truly remarkable paper [46] which laid down the foundations of the subject. We begin with a quote from this paper which a nutshell summarizes one of the main concerns of information theory:

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*

This paper proposed the definition of the capacity of a classical channel as the amount of information per channel use that can be transmitted asymptotically in the limit of many channel uses, with near perfect reproduction at the receiver's end, and gave a simple and elegant formula for the capacity. Here, the information is the logarithm (base 2) of the number of messages, in other words the number of classical bits that can be transmitted by the channel.

The definition of quantum channel capacity is motivated largely by the same problem, with the difference being that either the method of reproduction or the message itself involves fundamentally quantum effects. For many years, information theorists

either ignored quantum effects or approximated them so as to make them susceptible to classical analysis; it was only in the last decade or so that the systematic study of quantum information theory began.

Shannon's original paper set forth two coding theorems which form the foundation of the field of information theory. The first is the source coding theorem, which gives a formula for how much a random information source can be compressed. The second is the channel coding theorem, which gives a formula for how much redundancy must be added to a message in order to accurately reproduce it after sending the information through a noisy channel.

3 Shannon theory

Shannon's 1948 paper [46] contained two theorems for which we give quantum analogs. The first of these is the *source coding* theorem, which gives a formula for how much a source emitting random signals can be compressed, while still permitting the original signals to be recovered with high probability. Shannon's source coding theorem states that n outputs of a source X can be compressed to length $nH(X) + o(n)$ bits, and restored to the original with high probability, where H is the entropy function. For a probability distribution with probabilities p_1, p_2, ..., p_n, the entropy H is

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

where information theorists generally take the logarithm base 2 (thus obtaining bits as the unit of information).

The second of these theorems is the *channel coding* theorem, which states that with high probability, n uses of a noisy channel N can communicate $Cn - o(n)$ bits reliably, where C is the channel capacity given by

$$C = \max_{p(X)} I(X; N(X)) \tag{2}$$

Here the maximum is taken over all probability distributions on inputs X to the channel, and $N(X)$ is the output of the channel given input X. The *mutual information* I between two random variables X and Y is defined as:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (3)\\ &= H(X) + H(Y) - H(X,Y), && (4) \end{aligned}$$

where $H(X,Y)$ is the entropy of the joint distribution of X and Y, and $H(Y|X)$ is the conditional entropy of Y, given X. That is, if the possible values of X are $\{X_i\}$, then the conditional entropy is

$$H(Y|X) = \sum_i \Pr(X = X_i)\, H(Y|X = X_i). \tag{5}$$

There is an efficient algorithm, the *Arimoto-Blahut algorithm*, for calculating the capacity (2) of a classical channel [2, 12, 15, 43].

When the formula for mutual information is extended to the quantum case, two generalizations have been found that both give capacities of a quantum channel, although these capacities differ in both the resources that the sender and receiver have available and the operations they are permitted to carry out. One of these formulae generalizes the expression (3) and the other the expression (4); these expressions are equal in the classical case.

For classical channels, there are a number of extra resources which one might imagine could increase their capacity. These include a feedback channel from the receiver to the sender and shared randomness between the sender and the receiver. It turns out that neither of these resources actually does increase the capacity of a classical channel. For quantum channels, however, the situation is different. In this case, one of the resources that we must consider is entanglement. An entangled pair of quantum states consists of two states which are non-classically correlated. To parties who share such a pair of states cannot use them to transmit information, but can use them to obtain a shared random variable. It turns out that shared entanglement between the sender and receiver can be used to increase the transmission capacity of a quantum channel. When the capacity of a quantum channel for transmitting quantum information is considered, things become even more complicated. In this case, a classical side channel can increase the capacity of a quantum channel to transmit quantum information, even though no quantum information can be transmitted by a classical channel.

4 Quantum mechanics

Before we can start talking about quantum information theory, I need to give a brief description of some of the fundamental principles of quantum mechanics. The first of these principles that we present is the *superposition principle*. In its most basic form, this principle says that if a quantum system can be in one of two distinguishable states $|\,x\rangle$ and $|\,y\rangle$, it can be in any state of the form $\alpha\,|\,x\rangle + \beta\,|\,y\rangle$, where α and β are complex numbers with $|\alpha|^2 + |\beta|^2 = 1$. Here $|\,\cdot\rangle$ is the *bra-ket* notation that physicists use for a quantum state; we will occasionally be using it in the rest of this paper. Recall that we assumed that $|\,x\rangle$ and $|\,y\rangle$ were distinguishable, so there must conceptually be some physical experiment which distinguishes them (this experiment need not be performable in practice). The principle says further that if we perform this experiment, we will observe the state $|\,x\rangle$ with probability $|\alpha|^2$ and $|\,y\rangle$ with probability $|\beta|^2$. Furthermore, after this experiment is performed, if state $|\,x\rangle$ (or $|\,y\rangle$) is observed the system will thereafter behave in the same way as it would have had it originally been in state $|\,x\rangle$ (or $|\,y\rangle$).

Mathematically, the superposition principle says that the states of a quantum system are the unit vectors of a complex vector space, that two orthogonal vectors are distinguishable, and that measurement projects the state onto one of an complete orthonormal set of basis vectors. In accordance with physics usage, we will represent quantum states by column vectors. The Dirac *bra-ket* notation denotes a column vector by $|\,v\rangle$ (a *ket*) and its Hermitian transpose (i.e., complex conjugate transpose) by $\langle v\,|$ (a *bra*). The inner product between two vectors, v and w, is denoted $\langle w|v\rangle = w^\dagger v$, here we define $X^\dagger$ (whether X is a vector or matrix) to be the Hermitian transpose of X.

Multiplying a quantum state vector by a complex phase factor (a unit complex number) does not change any properties of the system, so mathematically the state of a quantum system is a point in projective complex space. Unless otherwise stated, however, we will represent quantum states as unit vectors in some complex vector space $\mathbb{C}^d$.

We will be dealing solely with finite dimensional vector spaces. For an introductory paper, quantum information theory is already complicated enough in finite dimensions without introducing the additional complexity of infinite-dimensional vector spaces. Many of the theorems we discuss do indeed generalize naturally to infinite-dimensional spaces.

A *qubit* is a two-dimensional quantum system. Probably the most widely known qubit is the polarization of a photon, and we will thus be using this example. For the polarization of a photon, there can only be two distinguishable states. If one sends a photon through a birefringent crystal, it will take one of two paths, depending on its polarization. By re-orienting this crystal, these two distinguishable polarization states can be chosen to be horizontal and vertical, or right diagonal and left diagonal. In accordance with the superposition principle, each of these states can be expressed as a complex combination of basis states in the other basis. For example,

$$\begin{aligned} |\nearrow\rangle &= \frac{1}{\sqrt{2}}|\leftrightarrow\rangle + \frac{1}{\sqrt{2}}|\updownarrow\rangle \\ |\nwarrow\rangle &= \frac{1}{\sqrt{2}}|\leftrightarrow\rangle - \frac{1}{\sqrt{2}}|\updownarrow\rangle \\ |\circlearrowright\rangle &= \frac{1}{\sqrt{2}}|\leftrightarrow\rangle + \frac{i}{\sqrt{2}}|\updownarrow\rangle \\ |\circlearrowleft\rangle &= \frac{1}{\sqrt{2}}|\leftrightarrow\rangle - \frac{i}{\sqrt{2}}|\updownarrow\rangle \end{aligned}$$

Here, $|\circlearrowright\rangle$ and $|\circlearrowleft\rangle$ stand for right and left circularly polarized light, respectively; these are another pair of basis states for the polarization of photons. For a specific example, when diagonally polarized photons are put through a birefringent crystal oriented in the $\updownarrow, \leftrightarrow$ direction, half of them will behave like vertically polarized photons, and half like horizontally polarized photons; thereafter, these photons will indeed have these polarizations.

If you have two quantum systems, their joint state space is the tensor product of their individual state spaces. For example, the state space of two qubits is $\mathbb{C}^4$ and of three qubits is $\mathbb{C}^8$. The high dimensionality of the space for n qubits, $\mathbb{C}^{2^n}$, is one of the places where quantum computation attains its power.

The polarization state space of two photons has as a basis the four states

$$|\updownarrow\updownarrow\rangle, \quad |\updownarrow\leftrightarrow\rangle, \quad |\leftrightarrow\updownarrow\rangle, \quad |\leftrightarrow\leftrightarrow\rangle.$$

This state space includes states such as an EPR (Einstein, Podolsky, Rosen) pair of photons

$$\frac{1}{\sqrt{2}}(|\updownarrow\leftrightarrow\rangle - |\leftrightarrow\updownarrow\rangle) = \frac{1}{\sqrt{2}}(|\nearrow\nwarrow\rangle - |\nwarrow\nearrow\rangle), \tag{6}$$

where neither qubit alone has a definite state, but which has a definite state when considered as a joint system of two qubits. In this state, the two photons have orthogonal

polarizations in whichever basis they are measured in. Bell [6] showed that the outcomes of measurements on the photons of this state cannot be reproduced by joint probability distributions which give probabilities for the outcomes of all possible measurements, and in which each of the single photons has a definite probability distribution for the outcome of measurements on it, independent of the measurements which are made on the other photon [6, 20]. In other words, there cannot be any set of hidden variables associated with each photon that determines the probability distribution obtained when this photon is measured in any particular basis. Two quantum systems such as an EPR pair which are non-classically correlated are said to be *entangled* [13].

The next fundamental principle of quantum mechanics we discuss is the *linearity principle.* This principle states that an isolated quantum system undergoes linear evolution. Because the quantum systems we are considering are finite dimensional vector spaces, a linear evolution of these can be described by multiplication by a matrix. It is fairly easy to check that in order to make the probabilities sum to one, we must restrict these matrices to be unitary (a matrix U is unitary if $U^\dagger = U^{-1}$; unitary matrices are those complex matrices which take unit vectors to unit vectors).

Although many elementary treatments of quantum mechanics restrict themselves to pure states (unit vectors), for quantum information theory we need to treat probability distributions over quantum states. These naturally give rise to objects called density matrices. For an n-dimensional quantum state space, a *density matrix* is an $n \times n$ Hermitian trace-one positive semidefinite matrix.

Density matrices arise naturally from quantum states in two ways. The first way in which density matrices arise is from probability distributions over quantum states. A rank one density matrix ρ corresponds to the pure state v where $\rho = vv^\dagger$. (Recall $v^\dagger$ was the Hermitian transpose of v.) Suppose that we have a system which is in state v_i with probability p_i. The corresponding density matrix is

$$\rho = \sum_i p_i v_i v_i^\dagger. \tag{7}$$

An important fact about density matrices is that the density matrix for a system gives as much information as it is possible to obtain about experiments performed on the system. That is, any two systems with the same density matrix ρ cannot be distinguished by experiments, provided that no extra side information is given about these systems.

The other way in which density matrices arise is through disregarding part of an entangled quantum state. Recall that two systems in an entangled pure state have a definite quantum state when considered jointly, but that neither of the two systems individually can be said to have a definite state. The state of either of these systems considered separately is naturally represented by a density matrix. Suppose that we have a state ρ_{AB} on a tensor product system $\mathcal{H}_A \otimes \mathcal{H}_B$. If we can only see the first part of the system, this part behaves as though it is in the state $\rho_A = \mathrm{Tr}_B \rho_{AB}$. Here, Tr_B is the partial trace operator. Consider a joint system in the state

$$\rho_{AB} = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix}. \tag{8}$$

In this example, the dimension of $\mathcal{H}_A$ is 3 and the dimension of $\mathcal{H}_B$ is the size of the matrices B_{ij}. The partial trace of ρ_{AB}, tracing over $\mathcal{H}_A$, is

$$\mathrm{Tr}_A\ \rho_{AB} = B_{11} + B_{22} + B_{33} \tag{9}$$

Although the above formula also determines the partial trace when we trace over $\mathcal{H}_B$, through a permutation of the coordinates, it is instructive to give this explicitly:

$$\mathrm{Tr}_B\ \rho_{AB} = \begin{pmatrix} \mathrm{Tr}\, B_{11} & \mathrm{Tr}\, B_{12} & \mathrm{Tr}\, B_{13} \\ \mathrm{Tr}\, B_{21} & \mathrm{Tr}\, B_{22} & \mathrm{Tr}\, B_{23} \\ \mathrm{Tr}\, B_{31} & \mathrm{Tr}\, B_{32} & \mathrm{Tr}\, B_{33} \end{pmatrix}. \tag{10}$$

The final ingredient we need before we can start explaining quantum information theory is a *von Neumann measurement*. We have seen examples of this process before, while explaining the superposition principle; however, we have not yet given the general mathematical formulation of a von Neumann measurement. Suppose that we have an n-dimensional quantum system $\mathcal{H}$. A von Neumann measurement corresponds to a complete set of orthogonal subspaces S_1, S_2, ..., S_k of $\mathcal{H}$. Here, complete means that the subspaces S_i span the space $\mathcal{H}$, so that $\sum_i \dim S_i = n$. Let Π_i be the projection matrix onto the subspace S_i. If we start with a density matrix ρ, the von Neumann measurement corresponding to the set of subspaces $\{S_i\}$ projects ρ into one of the subspaces S_i. Specifically, it projects ρ onto the i'th subspace with probability $\mathrm{Tr}\ \Pi_i\rho$, the state after the projection being

$$\frac{1}{\mathrm{Tr}\,\Pi_i\rho}\Pi_i\rho\Pi_i,$$

where we have renormalized the projection to have trace 1. A special case that is often encountered is when the S_i are all one-dimensional, so that $S_i = w_i w_i^\dagger$, and the vectors w_i form an orthogonal basis of $\mathcal{H}$. Then, a vector v is taken to w_i with probability $|w_i^\dagger v|^2$, and a density matrix ρ is taken to $w_i w_i^\dagger$ with probability $w_i^\dagger \rho w_i$.

5 Von Neumann entropy

We are now ready to consider quantum information theory. We will start by defining the entropy of a quantum system. To give some intuition for this definition, we first consider some special cases. Consider n photons, each being in the state $|\updownarrow\rangle$ or $|\leftrightarrow\rangle$ with probability $\frac{1}{2}$. Any two of these states are completely distinguishable. There are thus 2^n equally probable states of the system, and the entropy is n bits. This is essentially a classical system.

Consider now n photons, each being in the state $|\updownarrow\rangle$ or $|\nearrow\rangle$ with probability $\frac{1}{2}$. These states are not completely distinguishable, so there are effectively considerably less than 2^n states, and the entropy should intuitively be less than n bits.

By thermodynamic arguments involving the increase in entropy associated with the work extracted from a system, von Neumann deduced that the *(von Neumann) entropy* of a quantum system with density matrix ρ should be

$$H_{\mathrm{vN}}(\rho) = -\mathrm{Tr}\rho\log\rho. \tag{11}$$

Recall that ρ is positive semidefinite, so that $-\mathrm{Tr}\rho\log\rho$ is well defined. If ρ is expressed in coordinates in which it is diagonal with eigenvalues λ_i, then in these coordinates $-\rho\log\rho$ is diagonal with eigenvalues $-\lambda_i\log\lambda_i$. We thus see that

$$H_{\mathrm{vN}}(\rho) = H_{\mathrm{Shan}}(\lambda_i), \tag{12}$$

so that the von Neumann entropy of a density matrix is the Shannon entropy of the eigenvalues. (Recall $\mathrm{Tr}\rho = 1$, so that $\sum_i \lambda_i = 1$.) This definition is easily seen to agree with the Shannon entropy in the classical case, where all the states are distinguishable.

6 Source coding

Von Neumann developed the above definition of entropy for thermodynamics. One can ask whether this is also the correct definition of entropy for information theory. We will first give the example of quantum source coding [30, 44], also called *Schumacher compression*, for which we will see that it is indeed the right definition. We consider a memoryless quantum source that at each time step emits the pure state v_i with probability p_i. We would like to encode this signal in as few qubits as possible, and send them to a receiver who will then be able to reconstruct the original state. Naturally, we will not be able to transmit the original state flawlessly. In fact, the receiver cannot even reconstruct the original state absolutely perfectly most of the time (this is the corresponding requirement in classical information theory). Unlike classical signals, quantum states are not completely distinguishable theoretically, so reconstructing the original state most of the time is too stringent a requirement. What we require is that the receiver be able to reconstruct a state which is almost completely indistinguishable from the original state nearly all the time. For this we need a measure of indistinguishability; we will use a measure called *fidelity*. Suppose that the original signal is a vector

$$u = v_1 \otimes v_2 \otimes \ldots \otimes v_n.$$

Then the fidelity between the signal u and the output ρ (which is in general a mixed state, i.e., a density matrix, on n qubits) is $F = u^\dagger \rho u$. The average fidelity is this fidelity F averaged over u. If the output is also a pure state v, the fidelity $F = u^\dagger v v^\dagger u = |u^\dagger v|^2$. If the input is a pure state, the fidelity measures the probability of success of a test which determines whether the output is the same as the input. If both the output state ρ_{out} and the input state ρ_{in} are mixed states, the fidelity is defined

$$\mathrm{Tr}\sqrt{\rho_{\mathrm{out}}^{1/2}\rho_{\mathrm{in}}\rho_{\mathrm{out}}^{1/2}},$$

an expression which, despite its appearance, is symmetric in ρ_{in} and ρ_{out} [29]. In the case where either ρ_{out} or ρ_{in} is pure, this is equivalent to the previous definition, and for mixed states it is a relatively simple expression which gives an upper bound on the probability of distinguishing these two states.

Before I can continue to sketch the proof of the quantum source coding theorem, I need to review the proof of the classical source coding theorem. Suppose we have a

memoryless source, i.e., a source X that at each time step emits the i'th signal type S_i with probability p_i, and where the probability distribution for each signal is independent of the previously emitted signals. The idea behind classical source coding is to show that with high probability, the source emits a *typical sequence.* Here a sequence of length n is defined to be *typical* if it contains approximately np_i copies of the signal S_i for every i.[1] The number of typical sequences is only $2^{nH(X)+o(n)}$. These can thus be coded in $nH(X) + o(n)$ bits.

The tool that we use to perform Schumacher compression is that of *typical subspaces*. Suppose that we have a density matrix $\rho \in \mathcal{H}$, where $\mathcal{H} = \mathbb{C}^k$, and we take the tensor product of n copies of ρ in the space $\mathcal{H}^n$, i.e., we take $\rho^{\otimes n} \in \mathbb{C}^{nk}$. There is a typical subspace associated with $\rho^{\otimes n}$. Let $\hat{v}_1$, $\hat{v}_2$, ..., $\hat{v}_k$ be the eigenvectors of ρ with associated eigenvalues λ_1, λ_2, ..., λ_k. Since $\mathrm{Tr}\rho = 1$, these λ_i form a probability distribution. Consider typical sequences of the eigenvectors $\hat{v}_i$, where λ_i is the probability of choosing $\hat{v}_i$. A typical sequence can be turned into a quantum state in $\mathcal{H}^{\otimes n}$ by taking the tensor products of its elements. That is, if a typical sequence is $\hat{v}_{i_1}$, $\hat{v}_{i_2}$, ..., $\hat{v}_{i_n}$, the corresponding quantum state is $w = \hat{v}_{i_1} \otimes \hat{v}_{i_2} \otimes \ldots \otimes \hat{v}_{i_n}$. The typical subspace $\mathcal{T}$ is the subspace spanned by typical sequences of the eigenvectors. The subspace $\mathcal{T}$ has dimension equal to the number of typical sequences, or $2^{H_{\mathrm{vN}}(\rho)n+o(n)}$.

We can now explain how to do Schumacher compression. Suppose we wish to compress a source emitting v_i with probability p_i. Let the typical subspace corresponding to $\rho^{\otimes n}$ be $\mathcal{T}$, where $\rho = \sum_i p_i v_i v_i^\dagger$ is the density matrix for the source, and where we are using a block length n for our compression scheme. We take the vector $u = v_{i_1} \otimes v_{i_2} \otimes \ldots \otimes v_{i_n}$ and make the von Neumann measurement that projects it into either $\mathcal{T}$ or $\mathcal{T}^\perp$. If u is projected onto $\mathcal{T}$, we send the results of this projection to the receiver; this can be done with $\log \dim \mathcal{T} = nH_{\mathrm{vN}}(\rho) + o(n)$ qubits. If u is projected onto $\mathcal{T}^\perp$, our compression algorithm has failed and we can send anything; this does not degrade the fidelity of our transmission much, because this is a low probability event.

Why did this work? We give a brief sketch of the proof. The main element of the proof is to show that the probability that we project u onto $\mathcal{T}$ approaches 1 as n goes to ∞. This probability is $u^\dagger \Pi_{\mathcal{T}} u$. If this probability were exactly 1, then u would necessarily be in $\mathcal{T}$, and we would have noiseless compression. If this probability is close to 1, then u is close to the subspace $\mathcal{T}$, and so u has high fidelity with the projected vector $\Pi_{\mathcal{T}} u$. Suppose the probability that the state u is projected onto $\mathcal{T}$ is $1 - \epsilon$. Then $u^\dagger \, \Pi_{\mathcal{T}} \, u = 1 - \epsilon$ and the fidelity between the original state u and the final state is $|\langle u | \Pi_{\mathcal{T}} \, u \rangle|^2 = (1 - \epsilon)^2$.

Now, recall that if two density matrices are equal, the outcomes of any experiments performed on them have the same probabilities. Thus, the probability that the source v_i with probabilities p_i is projected onto the typical subspace is the same as for the source $\hat{v}_i$ with probabilities λ_i, where $\hat{v}_i$ and λ_i are the eigenvalues and eigenvectors of $\rho = \sum_i p_i v_i v_i^\dagger$. Because the $\hat{v}_i$ are distinguishable, this is essentially the classical case, and w is in the typical subspace exactly when the sequence of $\hat{v}_i$ is a typical sequence. We then know from the classical theory of typical sequences that $w = \hat{v}_{i_1} \otimes \hat{v}_{i_2} \otimes \ldots \otimes \hat{v}_{i_k}$ is in the typical subspace at least $1 - \epsilon$ of the time, completing the proof.

[1] Strictly speaking, this is the definition of frequency-typical sequences. Their is a related but distinct definition of entropy-typical sequences and subspaces, which can also be used in many of these proofs.

7 Accessible information and the $C_{1,1}$ capacity

The next concept we consider is that of *accessible information*. Here, we again have a source emitting state σ_i with probability p_i. Note that now, the states σ_i emitted may be density matrices rather than pure states. We will ask a different question this time. We now want to obtain as much information as possible about the sequence of signals emitted by the source. This is called the *accessible information* of the source. That is, the accessible information is the maximum over all measurements of the mutual information $I(X;Y)$ where X is the random variable telling which signal σ_i was emitted by the source, and Y is the random variable giving the outcome of a measurement on σ_i. This gives the capacity of a channel where at each time step the sender must choose one of the states σ_i to send, and must furthermore choose σ_i a fraction p_i of the time; and where the receiver must choose a fixed measurement that he will make on every signal received.

To find the accessible information, we need to maximize over all measurements. For this, we need to be able to characterize all possible quantum measurements. It turns out that von Neumann measurements are not the most general class of quantum measurements; the most general measurement is called a *positive operator valued measure*, or *POVM*. One way to describe these is as von Neumann measurements on a quantum space larger than the original space; that is, by supplementing the quantum state space by an *ancilla* space and taking a von Neumann measurement on the joint state space.

We now give a more effective, but equivalent, characterization of POVM's. For simplicity, we restrict our discussion to POVM's with a finite number of distinct outcomes; these turn out to be sufficient for studying capacities of finite dimensional channels. A POVM can be defined by a set of positive semidefinite matrices E_i satisfying $\sum_i E_i = I$. If a quantum system has matrix ρ, then the probability of the i'th outcome is

$$p_i = \mathrm{Tr}(E_i \rho) \tag{13}$$

For a von Neumann measurement, we take $E_i = \Pi_{S_i}$, the projection matrix onto the i'th orthogonal subspace S_i. The condition $\sum_i \Pi_{S_i} = I$ is equivalent to the requirement that the S_i are orthogonal and span the whole state space. To obtain the maximum information from a POVM, we can assume that the E_i's are pure states; if there is an E_i that is not rank one, then we can always achieve at least as much accessible information by refining that E_i into a sum $E_i = \sum_j E_{ij}$ where the E_{ij} are rank one.

We now give some examples of the measurements maximizing accessible information. The first is one of the simplest examples. Suppose that we have just two pure states in our ensemble, with probability $\frac{1}{2}$ each. For example, we could take the states $|\updownarrow\rangle$ and $|\nearrow\rangle$. Let us take $v_1 = (1,0)$ and $v_2 = (\cos\theta, \sin\theta)$. We will not prove it here, but the optimal measurement for these is the von Neumann measurement with two orthogonal vectors symmetric around v_1 and v_2. That is, the measurement with projectors

$$w_1 = \left(\cos(\tfrac{\pi}{2} + \tfrac{\theta}{2}), \sin(\tfrac{\pi}{2} + \tfrac{\theta}{2})\right) \tag{14}$$

$$w_2 = \left(\cos(-\tfrac{\pi}{2} + \tfrac{\theta}{2}), \sin(-\tfrac{\pi}{2} + \tfrac{\theta}{2})\right) \tag{15}$$

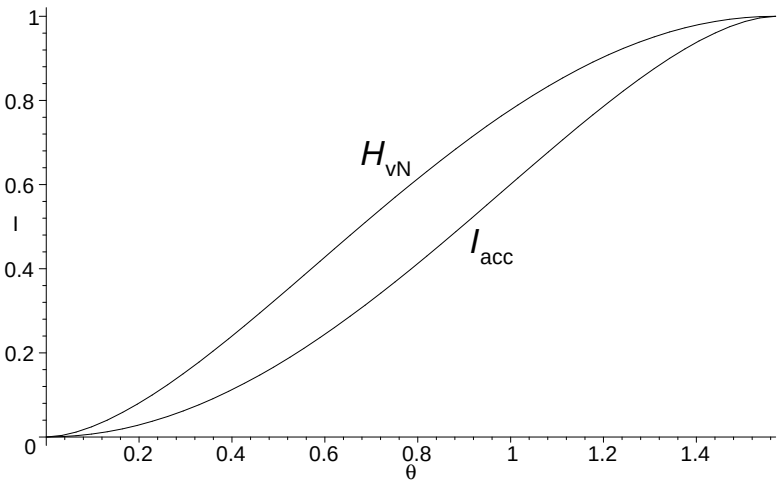

Figure 1: A plot of the von Neumann entropy of the density matrix and the accessible information for the ensemble of two pure quantum states with equal probabilities and that differ by an angle of θ, for $0 \leq \theta \leq \pi/2$. The top curve is the von Neumann entropy and the bottom the accessible information.

This measurement is symmetric with respect to interchanging v_1 and v_2, and it leads to a binary symmetric channel with error probability

$$\cos^2\left(\frac{\pi}{2}+\frac{\theta}{2}\right) = \frac{1}{2} - \frac{\sin\theta}{2}. \tag{16}$$

The accessible information is thus $1 - H_2(\frac{1}{2} - \frac{\sin\theta}{2})$. Here H_2 is the Shannon entropy of a binary signal, i.e.,

$$H_2(p) = -p\log p - (1-p)\log(1-p)$$

For the ensemble containing v_1 and v_2 with probability $\frac{1}{2}$ each, the density matrix is

$$\rho = \frac{1}{2}\begin{pmatrix} 1+\cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & 1-\cos^2\theta \end{pmatrix}, \tag{17}$$

which has eigenvalues $\frac{1}{2} \pm \cos\theta$, so the von Neumann entropy of the density matrix is $H_2(\frac{1}{2} - \frac{\cos\theta}{2})$. The values of $I_{\rm acc}$ and $H_{\rm vN}$ are plotted in Figure 1. One can see that the von Neumann entropy is larger than the accessible information.

Note that in our first example, the optimum measurement was a von Neumann measurement. If there are only two states in an ensemble, it has been conjectured that the measurement optimizing accessible information is always a von Neumann measurement, in part because extensive computer experiments have not found a counterexample [17]. This conjecture has been proven for quantum states in two dimensions [33]. Our next example shows that this conjecture does not hold for ensembles composed of three or more states.

Our second example is three photons with polarizations that differ by 60° each. These are represented by the vectors

$$\begin{aligned} v_0 &= (1,0) \\ v_1 &= (-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}) \\ v_2 &= (-\tfrac{1}{2}, -\tfrac{\sqrt{3}}{2}) \end{aligned}$$

The optimal measurement for these states is the POVM corresponding to the three vectors w_i where $w_i \perp v_i$. We take $E_i = \frac{2}{3} w_i w_i^\dagger$, in order for $\sum_i E_i = I$. If we start with vector v_i, it is easy to see that we never obtain w_i, but do obtain the other two possible outcomes with probability $\frac{1}{2}$ each. This gives an accessible information of $I_{\rm acc} = \log 3 - 1$. For these three signal states, it is also easy to check that the density matrix $\rho = \frac{1}{2} I$, so $H_{\rm vN} = 1$. Again, we have $I_{\rm acc} < H_{\rm vN}$.

Given these two examples and some intuition, one might formulate the conjecture that $I_{\rm acc} \leq H_{\rm vN}$. This is true, as in fact is a somewhat stronger theorem which we will shortly state. The first published proof of this theorem was given by Holevo [23]. It was earlier conjectured by Gordon [18] and stated by Levitin with no proof [32].

Theorem (Holevo): *Suppose that we have a memoryless source emitting an ensemble of (possibly mixed) states σ_i, where σ_i is emitted with probability p_i. Let*

$$\chi = H_{\rm vN}(\sum_i p_i \sigma_i) - \sum_i p_i H_{\rm vN}(\sigma_i). \tag{18}$$

Then

$$I_{\rm acc} \leq \chi. \tag{19}$$

The conditions for equality in this result are known. If all the σ_i commute, then they are simultaneously diagonalizable, and the situation is essentially classical. In this case, $I_{\rm acc} = \chi$; otherwise $I_{\rm acc} < \chi$.

We define the $C_{1,1}$ capacity of a quantum channel as the maximum over all ensembles of input states of the accessible information contained by the corresponding ensemble of output states. This is the capacity of the channel for transmitting quantum information if we restrict the protocols that we use; namely, we only allow protocols that do not send any states that are entangled over more than one channel use (this is the significance of the first '1' in the subscript), and do not perform any joint quantum measurements involving more than one channel output (this is the significance of the second '1'), and further we do not allow adaptive measurements of the outputs (i.e., the measurement chosen cannot depend on results of previous measurements on channel outputs). Allowing adaptive measurements can in certain circumstances increase the capacity, but they do not generally allow one reach the $C_{1,\infty}$ capacity discussed in the next section [48].

For example, if we consider the quantum channel where the sender can choose to convey to the receiver either of the two pure quantum states of our first example, the optimum ensemble is the ensemble consisting of both states with equal probability, and the $C_{1,1}$ capacity is $1 - H(\frac{1}{2} - \frac{\sin\theta}{2})$. For the channel which can convey to the

receiver any of the three states in our second example, the maximum ensemble giving $C_{1,1}$ turns out to be that which uses just two of the three states, each with probability $\frac{1}{2}$. This is our first example with $\theta = 60^\circ$, so this channel has a $C_{1,1}$ capacity of $1 - H(\frac{1}{2} - \frac{\sqrt{3}}{4}) \approx .6454$.

8 The classical capacity of a quantum channel

One can ask the question: is the $C_{1,1}$ capacity the most information that can be sent per quantum state, using only the three states of our second example? The answer is, surprisingly, "no". Suppose that we use the three length-two codewords $v_0 \otimes v_0$, $v_1 \otimes v_1$, and $v_2 \otimes v_2$. These are three pure states in the four-dimensional quantum space of two qubits. Since there are only three vectors, they lie in a three-dimensional subspace. The inner product between any two of these states is $\frac{1}{4}$. One can show for this tensor product ensemble, the optimal accessible information is attained by using the von Neumann measurement having three basis vectors obtained by "pulling" the three vectors $v_i \otimes v_i$ apart until they are all orthogonal. This measurement gives $I_{\rm acc} = 1.369$ bits, larger than twice the $C_{1,1}$ capacity, which is 1.2908 bits. We thus find that block coding and joint measurements let us achieve a better information transmission rate than $C_{1,1}$.

Having found that length two codewords work better than length one codewords, the natural question becomes: as the lengths of our codewords go to infinity, how well can we do. We define the $C_{1,\infty}$ capacity of a quantum channel as the capacity over protocols which do not permit inputs entangled between two or more channel uses, but do allow joint quantum measurements over arbitrarily many channel uses. A generalization of Shannon's giving the $C_{1,\infty}$ capacity has been proven.

Theorem (Holevo[24], Schumacher–Westmoreland[45]): *The $C_{1,\infty}$ capacity of a quantum channel, i.e., that capacity obtainable using codewords composed of signal states σ_i, where the probability of using σ_i is p_i, is*

$$\chi = H_{\rm vN}(\sum_i p_i \sigma_i) - \sum_i p_i H_{\rm vN}(\sigma_i). \tag{20}$$

Note that χ is a function of the probabilistic ensemble of the signal states $\{\sigma_i, p_i\}_i$ that we have chosen, where state σ_i has p_i. We will sometimes write $\chi(\{\sigma_i, p_i\}_i)$ so as to explicitly show this dependence. Another approach to proving this theorem, which also provides some additional results, appears in [37, 39]

We later give a sketch of the proof of the $C_{1,\infty}$ capacity formula in the special case where the σ_i are pure states. We will first ask: Does this formula give the true capacity of a quantum channel $\mathcal{N}$? There are certainly protocols in which the sender, for example, uses the two halves of an EPR pair of entangled qubits (as in Eq. (6)) as inputs for two separate channel uses. The question is: does allowing this type of protocol let one obtain a larger capacity?

Before we address this question (we will not be able to answer it) we should give the mathematical description of a general quantum channel. If $\mathcal{N}$ is a memoryless quantum communication channel, then it must take density matrices to density matrices. This means $\mathcal{N}$ must be a linear trace preserving positive map. Here, linear is required by

the basic principles of quantum mechanics and trace preserving is required since the channel must preserve trace 1 matrices. Positive means the channel takes positive semidefinite matrices to positive semidefinite matrices (and thus takes density matrices to density matrices). For $\mathcal{N}$ to be a valid quantum map, it must have one more property: namely, it must be completely positive. This means that $\mathcal{N}$ is positive even when it is tensored with the identity map. There is a theorem [22] that any linear completely positive map can be expressed as

$$\mathcal{N}(\rho) = \sum_i A_i \rho A_i^\dagger, \tag{21}$$

and the condition for the map to be trace-preserving is that the matrices A_i satisfy $\sum_i A_i^\dagger A_i = I$.

A natural guess at the capacity of a quantum channel $\mathcal{N}$ would be the maximum of χ over all possible distributions of channel outputs, that is,

$$\chi_{\max}(\mathcal{N}) = \max_{\{\sigma_i, p_i\}_i} \chi\big(\{\mathcal{N}(\sigma_i), p_i\}_i\big), \tag{22}$$

since the sender can effectively communicate to the receiver any of the states $\mathcal{N}(\sigma_i)$. We do not know whether this is the capacity of a quantum channel; if the use of entanglement between separate inputs to the channel helps to increase channel capacity, it would be possible to exceed this $\chi_{\max}$. This can be addressed by answering a question that is simple to state: Is $\chi_{\max}$ additive [1, 40, 31, 49, 36, 3]? That is, if we have two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$, is

$$\chi_{\max}(\mathcal{N}_1 \otimes \mathcal{N}_2) = \chi_{\max}(\mathcal{N}_1) + \chi_{\max}(\mathcal{N}_2). \tag{23}$$

Proving superadditivity of the quantity $\chi_{\max}$ (i.e., the $\geq$ direction of Eq. (23)) is easy. The open question is whether strictly more capacity can be attained by using the tensor product of two channels jointly than by using them separately.

We now return to the discussion of the proof of the Holevo-Schumacher-Westmoreland theorem in the special case where the σ_i are pure states. The proof of this case in fact appeared before the general theorem was proved [21]. The proof uses three ingredients. These are (1) random codes, (2) typical subspaces, and (3) the square root measurement.

The square root measurement is also called the "pretty good" measurement, and we have already seen an example of it. Recall our second example for accessible information, where we took the three vectors $v_i \otimes v_i$, where $v_i = (\cos\frac{2\pi i}{3}, \sin\frac{2\pi i}{3})$ for $i = 0, 1, 2$. The optimal measurement for I_{acc} on these vectors was the von Neumann measurement obtained by "pulling" them farther apart until they were orthogonal. This is an example of the square root measurement.

Suppose that we are trying to distinguish between vectors w_1, w_2, ..., w_n, which appear with equal probability (the square root measurement can also be defined for vectors having unequal probabilities, but we do not need this case). Let $\phi = \sum_i w_i w_i^\dagger$. The square root measurement has POVM elements $E_i = \phi^{-1/2} w_i w_i^\dagger \phi^{-1/2}$. We have

$$\sum_i E_i = \phi^{-1/2} \left(\sum_i w_i w_i^\dagger \right) \phi^{-1/2} = I, \tag{24}$$

so these E_i do indeed form a POVM.

We can now give the coding algorithm for the capacity theorem for pure states. We choose M codewords $u_j = v_{i_1} \otimes v_{i_2} \otimes \cdots \otimes v_{i_n}$, where the v_i are chosen at random with probability p_i. We then use these particular M codewords u_j to send information, where the coding scheme is chosen so that each of these codewords is sent with probability $\frac{1}{M}$. The difficult part of the proof is now showing that a random codeword can be identified with high probability.

To decode, we perform the following steps:

1. Project into the typical subspace $\mathcal{T}$. Most of the time, this projection works, and we obtain $w_j = (u^\dagger \Pi_{\mathcal{T}} u_j)^{-1/2} \Pi_{\mathcal{T}} u_j$, where $\Pi_{\mathcal{T}}$ is the projection matrix onto the subspace $\mathcal{T}$.

2. Use the square root measurement on the w_j.

The probability of error, given that the original state was w_j, is

$$\begin{aligned} 1 - w_j^\dagger E_j w_j &= 1 - w_j^\dagger \phi^{-1/2} w_j w_j^\dagger \phi^{-1/2} w_j \\ &= 1 - w_j^\dagger \phi^{-1/2} w_j \end{aligned}$$

The overall probability of error is thus

$$1 - \frac{1}{M} \sum_{j=1}^{M} |w_j^\dagger \phi^{-1/2} w_j|^2. \tag{25}$$

The intuition for why this procedure works (this intuition is apparently not even close to being rigorous, as the proof works along substantially different lines) is that for this probability of error to be small, we need that $\phi^{-1/2} w_j$ is close to w_j for most j. However, the w_j are distributed more or less randomly in the typical subspace $\mathcal{T}$, so $\phi = \sum_j w_j w_j^\dagger$ is moderately close to the identity matrix on its support, and thus $\phi^{-1/2} w_j$ is close to w_j. Note that we need that the number M of u_j is less than $\dim \mathcal{T}$, or otherwise it would be impossible to distinguish the w_j; as by Holevo's bound (19) a d-dimensional quantum state space can carry at most d bits of information.

9 Calculating the $C_{1,\infty}$ capacity.

We now consider the problem of numerically finding the $C_{1,\infty}$ capacity. Recall that this capacity was expressible as

$$\max_{\{p_i, v_i\}_i} H_{\text{vN}}(\mathcal{N}(\sum_i p_i v_i v_i^\dagger)) - \sum_i p_i H_{\text{vN}}(\mathcal{N}(v_i v_i^\dagger))$$

where the maximum is taken over all ensembles $\{p_i, v_i\}_i$ of pure states in the input space of the channel. We propose to maximize this in stages, at first holding one parameter of the ensemble fixed, and then holding other parameters of this ensemble fixed.

We first consider the problem of maximizing $C_{1,\infty}$ while holding the average density matrix $\rho = \sum_i p_i v_i v_i^\dagger$ fixed. This is equivalent to the minimization problem

$$\begin{aligned} \text{minimize} \qquad & \sum_i p_i H_{\text{vN}}(\mathcal{N}(v_i v_i^\dagger)) \\ \text{subject to} \qquad & \sum_i p_i v_i v_i^\dagger = \rho. \end{aligned} \tag{26}$$

The problem of finding the probability distribution p_i minimizing the expression (26 is a linear programming problem, albeit one with an infinite number of variables p_i, one for each of the continuum of pure states v_i. There is a standard way of attacking such problems called column generation. First, the linear program is solved with v_i restricted to be chosen from some fixed finite set of possible signal states (these correspond to columns of the linear program). We then find vectors v which, if added to this linear program, would yield a better solution. By iterating the steps of finding new vectors to add to the linear program and solving the resulting improved linear program, we hope to eventually converge upon the right solution. If we are guaranteed to find a good vector v_i to add if one exists, then it turns out that we will indeed converge upon the optimal solution.

We now give a few more details of this process. If the vectors v_i range over a d-dimensional input space, there are d^2 constraints in this problem, arising from the degrees of freedom of the matrix equality

$$\sum_i p_i v_i v_i^\dagger = \rho.$$

(Both $v_i v_i^\dagger$ and ρ are $d \times d$ Hermitian matrices, yielding d^2 degrees of freedom. Note that $\sum_i p_i = 1$ is implicit in this matrix equality, as this can be obtained by taking the trace of both sides.) There is thus an optimum solution which has at most d^2 non-zero values of p_i, one for each constraint of the problem.

The success of column generation is dependent on how well we can find a column which will be advantageous to add to the linear program. To describe how to do this, we need to introduce the concept of the dual of a linear program.

This dual is another linear program. The constraints of the first program (called the primal program) correspond to the variables of the dual program, and vice versa. If the primal program is a minimization, then the dual program is a maximization. The fundamental theorem of linear programming asserts that the optimum value of these two programs are equal.

We now give the dual of the program (26) above. The variables of this dual will be the entries of a Hermitian matrix τ. The program is

$$\begin{aligned} \text{maximize} \qquad & \operatorname{Tr} \tau\rho \\ \text{subject to} \qquad & v^\dagger \tau v \le H_{\text{vN}}(\mathcal{N}(vv^\dagger)) \quad \text{for all unit vectors } v \in \mathbb{C}^d. \end{aligned} \tag{27}$$

It may be instructive to consider the proof that the optimal value of the primal is greater than or equal to the optimal value of the dual. Suppose we have a τ such that $\operatorname{Tr} \tau\rho = \mathbf{x}$.

Then

$$\begin{aligned}\sum_i p_i H_{\rm vN}(\mathcal{N}(v_i v_i^\dagger)) &\geq \sum_i p_i v_i^\dagger \tau v_i \\ &= \sum_i p_i {\rm Tr}\, \tau v_i v_i^\dagger \\ &= {\rm Tr}\, \tau\rho \;=\; \mathbf{x}. \end{aligned} \tag{28}$$

Suppose we have an optimal solution of the primal program restricted to a given fixed set of variables v_i. This solution will correspond to a dual problem with some optimal Hermitian matrix τ. To find a good column to add to the primal, we need to find a v that violates the constraints of the dual, i.e., such that

$$H_{\rm vN}(\mathcal{N}(vv^\dagger)) - v^\dagger \tau v < 0 \tag{29}$$

This is a nonlinear optimization problem which I believe should be solvable in low dimensions by gradient descent. One can look for good vectors to add by starting at an arbitrary vector v and proceeding downhill to find local minima of the expression (29). One should start both at random points, and at the vectors v_i having nonzero probability in the current solution (these latter will improve the solution by adapting to the perturbation made to the problem since the last iteration). For high dimensions, this technique will grow inefficient exponentially fast, but the curse of dimensionality also may mean that the problem is intrinsically difficult.

We now need to show how to change the average density matrix ρ of our ensemble so as to improve the optimal value. Recall that we want to maximize

$$H_{\rm vN}(\mathcal{N}(\rho)) - \sum_i p_i H_{\rm vN}(\mathcal{N}(v_i v_i^\dagger)). \tag{30}$$

By the linear programming duality (28), this is smaller than

$$H_{\rm vN}(\mathcal{N}(\rho)) - {\rm Tr}\, \rho\tau, \tag{31}$$

for an arbitrary ensemble $\{p_i, v_i\}_i$, and equal at the current maximum. From the concavity of entropy, the expression (31) is concave in ρ. Thus, if there is no direction to change ρ that will increase the maximum (31), there is also no direction that will increase (30), and we have found the optimal value of ρ (at least for our current set of v_i). If there is a direction that increases (31), then we can use binary search to find the optimum distance to move ρ in that direction. For the complete linear program with a continuum of variables v_i, we can use a smoothness argument to show that this same direction will also increase the objective function. This argument, unfortunately, does not appear to carry over to the finite dimensional linear program on a fixed set of v_i that we actually are solving. If things work well, it may turn out that attempting to move in this direction will result in a procedure that always converges to the optimum. Otherwise, we may have to use the polyhedral structure of the solution to our finite dimensional linear program to discover a good direction to move ρ.

The derivatives for Eqs. (29) and (31) can both be calculated explicitly. This can be done using the estimate

$$H(\rho + \epsilon\Delta) = H(\rho - \epsilon {\rm Tr}\Delta \log\rho + O(\epsilon^2)$$

which holds for matrices with $\mathrm{Tr}\rho = 1$ and $\mathrm{Tr}\Delta = 0$. It can be derived from the integral expression for $\log(\rho + \epsilon\delta)$ given in Eq. (20) of [42].

We thus propose to find the value of $C_{1,\infty}$ numerically by iterating the following steps

1. With a fixed set of v_i, and the constraint $\sum_i p_i v_i v_i^\dagger = \rho$, solve the linear program (26).
2. Change ρ so as to improve the value of the linear program solution above.
3. Find a quantum states v corresponding to columns it would be advantageous to add to the linear program.

If none of the steps can improve the solution, then we have discovered the optimum value of $C_{1,\infty}$. The hard step is (3); this is a non-linear optimization problem for which we have no good criteria to test whether we have discovered the global optimum. In large dimensions, this will clearly be the bottleneck step in making the procedure impractical.

10 Calculating the $C_{1,1}$ capacity.

We now very briefly describe how the ideas of the last section might be used to give a heuristic procedure for finding the $C_{1,1}$ capacity. Our technique again uses column generation. This time we propose to alternate between optimizing the ensemble of signal states used for the input, and optimizing the measurement used by the receiver. This is not guaranteed to find the global optimum; the second example of section 7 has four different local optima that are stable points in this procedure; the three ensembles each containing two of the signal states with probability $\frac{1}{2}$, and the ensemble containing all three with probability $\frac{1}{3}$ [48].

To find the optimal signal states, given a fixed measurement, we can in fact use exactly the procedure given in the previous section. By fixing the measurements, we have defined a quantum channel, with the input being a quantum state and the output being the results of the measurement, and so the procedure of the previous section is applicable. This is not an arbitrary quantum channel, as the output state is classical. Unfortunately, at this point we do not see how to use this fact to simplify the calculation of the capacity.

Finding the optimal measurement, given the signal states, can be done using essentially the same ideas as in the previous section. The measurement we use must extract a maximal amount of information, and so we can assume that each E_i takes the form $E_i = q_i w_i w_i^\dagger$ for some unit quantum state vector w_i in the output space. The condition that these form a POVM is

$$\sum_i q_i w_i w_i^\dagger = I.$$

Now, the expression for the capacity is the entropy of the input minus the entropy of the input, given the output. This can be seen to be linear in q_i. Again, we have an infinite dimensional linear program, which we can solve using the technique of

column generation. In this case, it is even slightly simpler; because the constraints are $\sum_i q_i w_i w_i^\dagger = I$, we do not need to incorporate any additional steps of optimizing ρ.

11 Entanglement-assisted capacities

In this section, we define the entanglement-assisted capacity of a quantum channel, and give the expression for it. For motivation, we first describe two surprising phenomena in quantum communication: *superdense coding* and *quantum teleportation.*

The process of *superdense coding* uses a shared EPR pair and a single qubit to encode two classical bits [11]. This is an improvement on the capacity of a noiseless, unassisted, quantum channel, which takes one quantum bit to send a classical bit. We will assume that the shared EPR pair is in the state

$$\frac{1}{\sqrt{2}}\left(|\,01\rangle - |\,10\rangle\right)$$

where the sender holds the first qubit and the receiver holds the second qubit. In this protocol, the sender starts by taking an EPR pair and applying to it either the identity operation or one of the three Pauli matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

He then sends his qubit to the receiver. The receiver now holds one of the four quantum states

$$\frac{1}{\sqrt{2}}\left(|\,01\rangle - |\,10\rangle\right), \qquad \frac{1}{\sqrt{2}}\left(|\,11\rangle - |\,00\rangle\right),$$
$$\frac{i}{\sqrt{2}}\left(|\,11\rangle + |\,00\rangle\right), \qquad \frac{1}{\sqrt{2}}\left(|\,01\rangle + |\,10\rangle\right).$$

These four states are known as the Bell basis, and they are mutually orthogonal. The receiver can thus uniquely identify which of these states he has, and so can unambiguously identify one of four messages, or two bits. (See Figure 2.)

There is a converse process to superdense coding known as quantum teleportation. It is impossible to send a quantum state over an unassisted classical channel. However, quantum teleportation lets a sender and a receiver who share an EPR pair of qubits communicate one qubit by sending two classical bits and using this EPR pair [8]. (See Figure 3.) In quantum teleportation, the sender measures the unknown quantum state state and the EPR pair in the Bell basis, and sends the receiver the two classical bits which are the results of this measurement. The receiver then performs a unitary operation. The measurements the sender makes are the same ones the receiver makes in superdense coding, and the unitary transformations the receiver performs are those the sender performs in superdense coding.

Quantum teleportation is a counterintuitive process, which at first sight seems to violate certain laws of physics; however, upon closer inspection one discovers that no

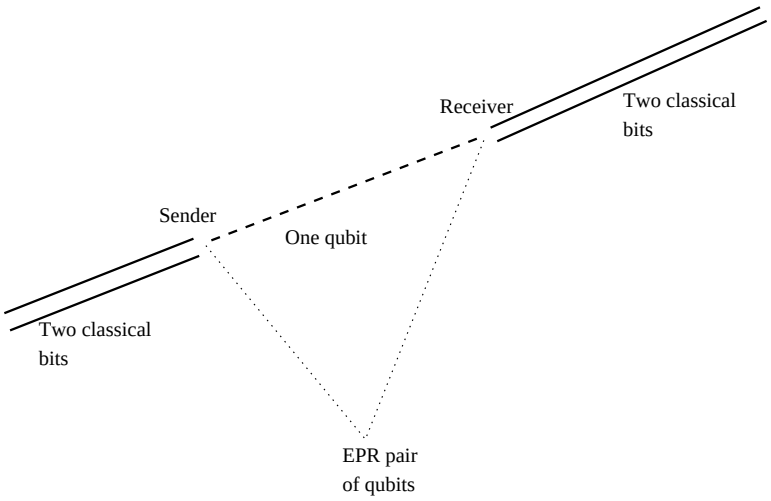

Figure 2: A schematic drawing of superdense coding. The sender can communicate two classical bits to the receiver using one qubit and a shared EPR pair. Here, the sender makes the same unitary transformation that the receiver would make in quantum teleportation, and the receiver makes the joint measurement that the sender would make in quantum teleportation.

actual paradoxes arise from teleportation. Teleportation cannot be used for superluminal communication, because the classical bits must travel at or slower than the speed of light. A continuous quantum state, which at first sight appears to contain many more than two bits of information, appears to have been transported using two discrete bits; however, by Holevo's bound, Eq. (19), one qubit can be used to transport at most one classical bit of information, so it is not possible to increase the capacity of a classical channel by encoding information in a teleported qubit. Finally, there is a theorem of quantum mechanics that an unknown quantum state cannot be duplicated [52]. However, this no-cloning theorem (as it is known) is not violated; the original state is necessarily destroyed by the measurement, so teleportation cannot be used to clone a quantum state.

We now give another capacity for quantum channels, one which has a capacity formula which can actually be completely proven, even in the case of infinite-dimensional Hilbert spaces [9, 10, 25, 26]. Recall that if $\mathcal{N}$ is a noiseless quantum channel, and if the sender and receiver possess shared EPR pairs, they can use superdense coding to double the classical information capacity of $\mathcal{N}$. Similarly, if $\mathcal{N}$ is a noiseless classical channel, EPR pairs can increase the capacity of the channel to send quantum information from zero qubits to half a qubit per channel use.

In general, if $\mathcal{N}$ is a noisy quantum channel, using shared EPR pairs can increase both the classical and quantum capacities of $\mathcal{N}$. With the aid of entanglement, the capacity for sending quantum information becomes exactly half of the capacity for classical information (this is a direct consequence of the phenomena of superdense

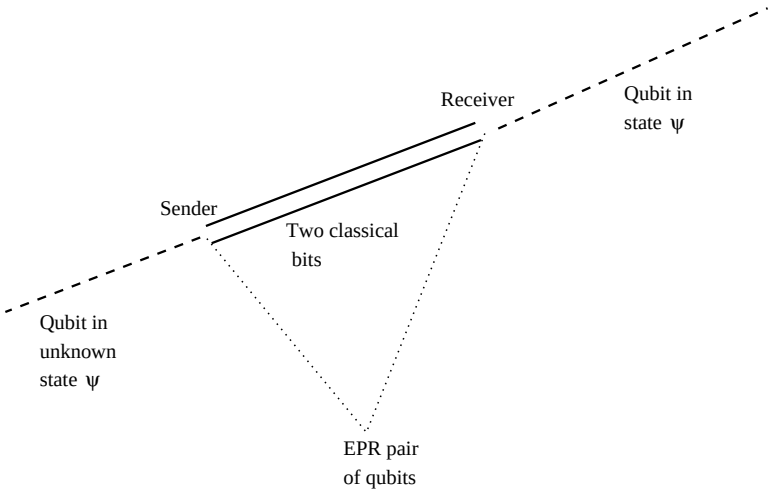

Figure 3: A schematic drawing of quantum teleportation. The sender has a qubit in an unknown state ψ that he wishes to send to the receiver. He also has half of an EPR state which he shares with the receiver. The sender makes a joint measurement on the unknown qubit and half of his EPR state, and communicates the results (2 classical bits) to the receiver. The receiver then makes one of four unitary transformations (depending on the two classical bits he received) on his half of the EPR state to obtain the state ψ.

coding and teleportation). We define the entanglement assisted capacity, C_E, as the quantity of classical information that can asymptotically be sent per channel use if the sender and receiver have access to a sufficient quantity of shared entanglement.

Theorem (Bennett, Shor, Smolin, Thapliyal [9, 10]): *The entanglement assisted capacity is*

$$C_E(\mathcal{N}) = \max_{\rho \in \mathcal{H}_{\rm in}} H_{\rm vN}(\rho) + H_{\rm vN}((\mathcal{N}(\rho)) - H_{\rm vN}((\mathcal{N} \otimes \mathcal{I})(\Phi_\rho)) \tag{32}$$

where $\rho \in \mathcal{H}_{\rm in}$ is a density matrix over the input space. Here, Φ_ρ is a pure state over the tensor product space $\mathcal{H}_{\rm in} \otimes \mathcal{H}_R$ such that $\mathrm{Tr}_R \Phi_\rho = \rho$. Here $\mathcal{H}_{\rm in}$ is the input state space and $\mathcal{H}_R$ is a reference system. The third term of the right hand side of (32), $H_{\rm vN}((\mathcal{N} \otimes \mathcal{I})(\Phi_\rho))$, is the entropy of the state resulting after the first half of Φ_ρ is sent through the channel $\mathcal{N}$, while the identity operation is applied to second half. The value of this term is independent of which reference system $\mathcal{H}_R$ and which pure state Φ_ρ are chosen.

The quantity being minimized in the above formula (32) is sometimes called quantum mutual information, and it is a generalization of the expression for mutual information in the form of Eq. (4). The proof of this result uses typical subspaces, superdense coding, the Holevo-Schumacher-Westmoreland theorem on the classical capacity of a quantum channel, and the strong subadditivity property of von Neumann entropy. The entanglement assisted capacity is a convex function of $\mathrm{Tr}_S \rho$, and so can be maximized with a straightforward application of gradient descent.

In the entanglement-assisted capacity, the communication protocol consumes the resource of entanglement. In general, it takes $H_{\rm vN}(\rho)$ bits of entanglement (i.e., EPR pairs) per channel use to achieve the capacity $C_E(\mathcal{N})$ in Eq. (32). I have also conjectured a formula for the capacity of a quantum channel using protocols which are only allowed to use a limited amount of entanglement. This formula is as follows.

Conjecture: *If the available entanglement per channel use is B bits, the capacity available is*

$$\max_{\substack{\rho_i, p_i \\ \sum_i p_i H(\rho_i) \leq B}} \sum_i p_i H_{\rm vN}(\rho_i) + H_{\rm vN}((\mathcal{N}(\sum_i p_i \rho_i)) - \sum_i p_i H_{\rm vN}((\mathcal{N} \otimes \mathcal{I})(\Phi_{\rho_i})). \quad (33)$$

where $\mathrm{Tr}_2 \Phi_{\rho_i} = \rho_i$, as in Eq. (32). Here, the maximization is over all probabilistic ensembles of density matrices $\{\rho_i, p_i\}_i$ where $\rho_i \in \mathcal{H}_{\rm in}$, $\sum_i p_i = 1$, and the average entropy of the ensemble, $\sum_i p_i H_{\rm vN}(\rho_i)$, is at most B.
I have a protocol which achieves this bound, and I can prove a matching upper bound over a restricted class of protocols.

A heuristic for finding this capacity with assistance by limited entanglement can be constructed using linear programming, column generation, and non-linear optimization, along the same lines as the protocol of Section 9. An extra constraint must be added that bounds the average entropy of the ensemble. One additional difficulty will be that the non-linear optimization problem needed to find good columns to add appears to become substantially harder.

12 Sending Quantum Information

Finally, we briefly mention the problem of sending quantum information (i.e., a quantum state) over a noisy quantum channel. In this scenario, several of the theorems that make classical channel capacity behave so nicely are demonstrably not true. Here, a feedback channel from the receiver to the sender, or a classical two-way side channel, will increase the quantum channel capacity, leading to several different capacities for transmitting quantum information. For the two-way quantum capacity, Q_2, the sender and receiver have a classical side channel they can use for free, For the quantum capacity with feedback, $Q_{FB} < Q_2$, the receiver has a classical feedback channel from himself to the sender. For the one-way quantum capacity, $Q \leq Q_{FB}$, all communication is directly from the sender to the receiver over the noisy quantum channel $\mathcal{N}$. The quantities Q_2 is closely related to a quantity defined on quantum states called the distillable entanglement [13]. Despite substantial study, not only do we not have any good ways to compute either Q_2 or Q_{FB}, but we also have no simple capacity formulas for representing it. There is a capacity formula for the one-way quantum capacity Q. It is essentially the last two terms of the expression (32) for entanglement-assisted capacity

$$Q(\mathcal{N}) = \lim_{n \to \infty} \frac{1}{n} \max_{\rho \in \mathcal{H}_{\rm in}} H_{\rm vN}(\mathcal{N}^{\otimes n}(\rho)) - H_{\rm vN}((\mathcal{N}^{\otimes n} \otimes \mathcal{I})\Phi_\rho) \quad (34)$$

where ρ, $\mathcal{H}_{\rm in}$ and Φ_ρ are defined as in (32). The quantity being maximized (before the limit $n \to \infty$) is called the *coherent information*. We need to take the maximum over the tensor product of n uses of the channel, and let n go to infinity, because unlike the

classical (or the quantum) mutual information, the coherent information is not additive [16]. The quantity Q(N) (34) is the quantum capacity of a noisy quantum channel $\mathcal{N}$ [34, 5, 27, 50]. Even for maximizing the single-symbol expression (that is, taking $n = 1$ in Eq. (34)), the calculation of the coherent information appears to be a difficult optimization problem, as there may be multiple local maxima. It would be a significant accomplishment to discover a good means of calculating this; unfortunately, I do not have any useful suggestions.

Acknowledgement

I would like to thank Mary Beth Ruskai for encouraging me to work on the question of calculating quantum channel capacities, as well as for her help with some the technical issues in quantum channel capacity.

References

[1] G. G. Amosov, A. S. Holevo, R. F. Werner, "On some additivity problems in quantum information theory," *Problems in information transmission*, vol. 36, pp. 25–34, 2000; LANL e-print math-ph/0003002, available at `http://xxx.lanl.gov`.

[2] S. Arimoto, "An algorithm for calculating the capacity of an arbitrary discrete memoryless channel," *IEEE Trans. Info. Theory*, vol. 18, pp. 14–20, 1972.

[3] K. M. R. Audenaert, and S. L. Braunstein, "On strong superadditivity of the entanglement of formation," LANL e-print quant-ph/0303045, available at `http://xxx.lanl.gov`.

[4] K. Audenaert, F. Verstraete, B. De Moor, "Variational characterizations of separability and entanglement of formation," *Phys. Rev. A*, vol. 64, art. 052304, 2001; LANL e-print quant-ph/0006128, available at `http://xxx.lanl.gov`.

[5] H. Barnum, J. A. Smolin and B. M. Terhal, "Quantum capacity is properly defined without encodings," *Phys. Rev. A*, vol. 58, pp. 3496–3501, 1998.

[6] J. S. Bell, "On the Einstein-Podolsky-Rosen paradox," *Physics*, vol. 1, pp. 195–200, 1964.

[7] C. H. Bennett, H. J. Bernstein, S. Popescu and B. Schumacher, "Concentrating partial entanglement by local operations," *Phys. Rev. A*, vol. 53, pp. 2046–2052, 1996.

[8] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," *Phys. Rev. Lett.*, vol. 70, pp. 1895–1899, 1993.

[9] C. H. Bennett, P. W. Shor, J. A. Smolin and A. V. Thapliyal, "Entanglement-assisted classical capacity of noisy quantum channels," *Phys. Rev. Lett.*, vol. 83, pp. 3081–3084, 1999.

[10] C. H. Bennett, P. W. Shor, J. A. Smolin and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem",

IEEE Trans. Info. Theory, vol. 48, pp. 2637–2655l 2002; LANL e-print quant-ph/0106052, available at http://xxx.lanl.gov.

[11] C. H. Bennett and S. J. Wiesner, "Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states," *Phys. Rev. Lett.*, vol. 69, pp. 2881–2884, 1992.

[12] R. E. Blahut "Computation of channel capacity and rate distortion functions," *IEEE Trans. Info. Theory*, vol. 18, pp. 460–473, 1972.

[13] D. Bruß, "Characterizing entanglement," *J. Math. Phys.* vol. 43, pp. 4237–4251, 2002. LANL e-print quant-ph/0110078, available at http://xxx.lanl.gov.

[14] V. Chvatal, *Linear Programming*, W. H. Freeman, New York, 1983.

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

[16] D. DiVincenzo, J. A. Smolin and P. W. Shor, "Quantum-channel capacity of very noisy channels, *Phys. Rev. A*, vol. 57, pp. 830–839, 1998.

[17] C. A. Fuchs and A. Peres, personal communication.

[18] J. P. Gordon, "Noise at optical frequencies; information theory," in *Quantum Electronics and Coherent Light; Proceedings of the International School of Physics Enrico Fermi, Course XXXI*, (P. A. Miles, ed., Academic Press New York, 1964) pp. 156–181.

[19] D. Gottesman, "An introduction to quantum error correction," in *Quantum Computation*, (S. J. Lomonaco, Jr., ed., AMS Press, Providence, RI, 2002); LANL e-print quant-ph/0004072, available at http://xxx.lanl.gov.

[20] D. M. Greenburger, M. A. Horne, A. Shimony and A. Zeilinger, "Bell's theorem without inequalities," *Am. J. Phys.*, vol. 58, pp. 1131-1143, 1990.

[21] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters, "Classical information capacity of a quantum channel," *Phys. Rev. A*, vol. 54, pp. 1869–1876, 1996.

[22] K. Hellwig and K. Krauss, "Operations and measurements II," *Communications in Mathematical Physics*, vol. 16, pp. 142-147, 1970.

[23] A. S. Holevo, "Information theoretical aspects of quantum measurements," *Probl. Info. Transm. (USSR)*, vol. 9, no. 2, pp. 31–42, 1973 (in Russian); [translation: A. S. Kholevo, *Probl. Info. Transm.*, vol. 9, pp. 177–183, 1973].

[24] A. S. Holevo, "The capacity of the quantum channel with general signal states," *IEEE Trans. Info. Theory*, vol. 44, pp. 269–273, 1998.

[25] A. S. Holevo, "On entanglement-assisted classical capacity," *J. Math. Phys.* vol. 43, pp. 4326–4333 (2002); LANL e-print quant-ph/0106075, available at http://xxx.lanl.gov

[26] A. S. Holevo, "Entanglement-assisted capacity of constrained channels," LANL e-print quant-ph/0211170, available at http://xxx.lanl.gov.

[27] M. Horodecki and S. Lloyd, manuscript in progress.

[28] M. Horodecki, P. Horodecki and R. Horodecki, "Unified approach to quantum capacities: Towards a quantum noisy coding theorem," LANL e-print quant-ph/0003040, available at http://xxx.lanl.gov.

[29] R. Jozsa, "Fidelity for mixed quantum states," J. Modern Optics, vol. 41, pp. 2315–2323, 1994.

[30] R. Jozsa and B. Schumacher, "A new proof of the quantum noiseless coding the-

orem," *J. Modern Optics*, vol. 41, pp. 2343-2349, 1994.

[31] C. King, "The capacity of the quantum depolarizing channel," *IEEE Trans. Inform. Theory*, vol.. 49, pp. 221–229, 2003; LANL e-print quant-ph/0204172, available at http://xxx.lanl.gov.

[32] L. B. Levitin, "On the quantum measure of information," in *Proceedings of the Fourth All-Union Conference on Information and Coding Theory, Sec. II*, Tashkent, 1969.

[33] L. B. Levitin, "Optimal quantum measurements for pure and mixed states," in *Quantum Communications and Measurement,* (V. P. Belavkin, O. Hirota and R. L. Hudson, eds., Plenum Press, New York and London, 1995) pp. 439–448.

[34] S. Lloyd, "The capacity of a noisy quantum channel," *Phys. Rev. A*,, vol. 55, pp. 1613-1622, 1997.

[35] H.-K. Lo and H. F. Chau, "Unconditional security of quantum key distribution over arbitrarily long distances," *Science*, vol. 283, pp. 2050–2056, 1999.

[36] K. Matsumoto, T. Shimono and A. Winter, "Remarks on additivity of the Holevo channel capacity and of the entanglement of formation, LANL e-print quant-ph/0206148, available at http://xxx.lanl.gov.

[37] M. Hayashi and H. Nagaoka, "General formulas for capacity of classical-quantum channels," LANL e-print quant-ph/0206186, available at http://xxx.lanl.gov.

[38] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 2000.

[39] T. Ogawa and H. Nagaoka, "A new proof of the channel coding theorem via hypothesis testing in quantum information theory," quant-ph 020813. LANL e-print quant-ph/0208139, available at http://xxx.lanl.gov.

[40] S. Osawa and H. Nagaoka, "Numerical experiments on the capacity of quantum channel with entangled input states," LANL e-print quant-ph/0007115, available at http://xxx.lanl.gov.

[41] J. R. Pierce, "The early days of information theory," *IEEE Trans. Info. Theory*, vol. 19, pp. 3–8, 1973.

[42] M. B. Ruskai, "Inequalities for quantum entropy: a review with conditions for equality" *J. Math. Phys.* vol. 43, pp. 4358-4375, 2002. LANL e-print quant-ph/0205064, available at http://xxx.lanl.gov.

[43] J. Sayir, "Iterating the Arimoto-Blahut algorithm for faster convergence," *Proceedings International Symposium on Information Theory*, p. 235, Sorrento, Italy, 2000.

[44] B. Schumacher, "Quantum coding," *Phys. Rev. A*, vol. 51, pp. 2738–2747, 1995.

[45] B. Schumacher and Westmoreland, "Sending classical information via a noisy quantum channel," *Phys. Rev. A*, vol. 56, pp. 131–138, 1997.

[46] C. E. Shannon, "A mathematical theory of communication," *The Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[47] P. W. Shor, "Quantum information theory: results and open problems," *Geom. Func. Anal. Special Volume–GAFA 2000* pp. 816–838 (2000).

[48] P. W. Shor, "The adaptive classical capacity of a quantum channel," LANL e-print quant-ph/0206058, available at http://xxx.lanl.gov.

[49] P. W. Shor, "Additivity of the classical capacity of entanglement-breaking chan-

nels," *J. Math. Physics*, vol. 43, pp. 4334-4340 (2002).
[50] P. W. Shor, manuscript in progress.
[51] R. R. Tucci, "Relaxation method for calculating quantum entanglement," LANL e-print quant-ph/0101123, available at `http://xxx.lanl.gov`.
[52] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, pp. 802–803, 1982.