Indeterminate-length quantum coding

Benjamin Schumacher
Department of Physics, Kenyon College, Gambier, Ohio 43022

Michael D. Westmoreland*
Department of Mathematical Sciences, Denison University, Granville, Ohio 43023
(Received 7 November 2000; published 10 September 2001)

The quantum analogues of classical variable-length codes are *indeterminate-length* quantum codes, in which code words may exist in superpositions of different lengths. This paper explores some of their properties. The length observable for such codes is governed by a quantum version of the Kraft–McMillan inequality. Indeterminate-length quantum codes also provide an alternate approach to quantum data compression.

DOI: 10.1103/PhysRevA.64.042304 PACS number(s): 03.67.−a, 03.65.Ta

I. INTRODUCTION

The development of the quantum information theory is a striking example of the fruitful hybridization of two well-established disciplines. Both quantum mechanics and information theory have a rich set of concepts and a powerful toolbox of mathematical techniques. Their combination is yielding powerful insights into the physical meaning of ''information'' [1,2].

One approach to this exploration is to begin with an idea of a ''classical'' information theory and investigate how this idea must be reinterpreted or modified to fit into the quantum information framework. Ideas of fidelity, quantum data compression [3], quantum error correcting codes [4], and the capacities of various quantum channels [5] can all be viewed in this light.

A basic idea in the classical theory of data compression is the idea of a variable-length code. A variable-length code assigns code words consisting of different numbers of symbols to different messages. If shorter code words are used for more common messages and longer ones for less common messages, the average code word length can be made shorter than would be possible using a fixed-length code. (Natural languages take advantage of this idea. Common words like ''the'' are often very short, while unusual words like ''sesquipedalian'' are longer.)

However, the original development of quantum data compression followed a different route, parallel to the classical development based on ''typical sequences.'' This left open the question of whether there was a quantum analog to classical variable-length coding. Because a quantum code must allow superpositions of different code words—including superpositions of code words of *different lengths*—the quantum version would best be termed an *indeterminate*-length quantum code.

One of us [6] made a preliminary investigation of this idea several years ago. Subsequently, Braunstein *et al.* [7] presented a quantum analog to classical Huffman coding. Because a general understanding of indeterminate-length quantum codes was not available then, Braunstein *et al.* were led to construct their code in an unnecessarily inefficient way. (See the discussion in Sec. II E.) More recently, Chuang and Modha have developed a quantum version of arithmetic coding as a route to quantum data compression [8]. Boström has also investigated indeterminate-length codes in connection with lossless quantum coding [9].

Our aim in this paper is to outline a general theory of indeterminate-length quantum codes, including their application to quantum data compression. We will first sketch a framework for discussing such codes. Each code will have a ''code word length'' observable Λ with integer eigenvalues; allowable code words include not only length eigenstates but arbitrary superpositions of them. The key requirement is that such codes be ''condensable''—that is, that the individual code words can be assembled into a string by means of a unitary operation. This condition leads us to prove a quantum version of the Kraft-McMillan inequality. Among the condensable codes are those that satisfy a quantum ''prefix-free'' condition, and we show (by giving an explicit condensation algorithm) that all such codes are condensable. We also show how classical variable-length codes can be used to construct quantum indeterminate-length codes with analogous properties.

We next turn to the use of indeterminate-length codes for quantum data compression. We achieve quantum data compression by taking a condensed string of N code words (in general, no shorter than N times the largest eigenvalue of Λ) and truncating it after the first $N\ell$ qubits, thus using only ℓ qubits per input code word. We show that the average $\langle l \rangle$ of the code word length observable Λ is the necessary and sufficient value of ℓ to achieve high fidelity for this process. It turns out that $\langle l \rangle$ is related to the quantum entropy S of the quantum information source, and from this relation we are able to arrive at the noiseless quantum coding theorem.

II. INDETERMINATE-LENGTH CODES

A. Zero-extended forms

In a quantum code, code words are states of finite strings of qubits. Superpositions of code words are also valid code words, and to maintain high fidelity we must preserve the coherence of these superpositions in our coding and decoding processes.

*Electronic address: westmoreland@denison.edu

 ©2001 The American Physical Society

We wish to create a code in which different code words have different *lengths*—that is, they involve different numbers of qubits. But how do we make sense of this idea? We will begin by considering *zero-extended forms* (ZEF) of the code words. For ZEF code words, we imagine that the code words are sitting at the beginning of a qubit register of fixed length, with $|0\rangle$'s following. These code words span a subspace of the Hilbert space of register states.

Our first essential requirement is that the code words carry their own length information. That is, we require that there is a ''length'' observable Λ on the ZEF code word subspace with the following two properties:

(a) The eigenvalues of Λ are $1,\ldots,l_{\max}$, where $l_{\max}$ is the length of the register, and (b) if $|\psi_{\mathrm{ZEF}}\rangle$ is an eigenstate of Λ with eigenvalue l, then it has the form

$$|\psi_{\mathrm{ZEF}}\rangle = |\psi^{1\cdots l}0^{l+1\cdots l_{\max}}\rangle. \tag{1}$$

In other words, the last $l_{\max}-l$ qubits in the register are in the state $|0\rangle$ for a ZEF code word of length l.

The length observable Λ was also considered in Ref. [5].

For each $l=1,\ldots,l_{\max}$, we let d_l be the dimension of the subspace spanned by the Λ eigenstates with eigenvalue l. Denote the projection onto this subspace by π_l. Then $\mathrm{Tr}\,\pi_l = d_l$.

B. Condensable codes

We want to be able to make use of the comparative shortness of some code words by ''packing'' the code words together, eliminating the trailing zeroes that ''pad'' the ends of the ZEF code words. But this must be a process that maintains quantum coherences in superpositions of code word states—that is, it must be described by a *unitary* transformation. Furthermore, we wish to be able to coherently pack together any number of code words.

We say that a code is *condensable* if the following condition holds: For any N, there is a unitary operator U (depending on N) that maps

$$\underbrace{|\psi_{1,\mathrm{ZEF}}\rangle\otimes\cdots\otimes|\psi_{N,\mathrm{ZEF}}\rangle}_{Nl_{\max}\ \text{qubits}} \rightarrow \underbrace{|\Psi_{1\cdots N,\mathrm{ZEF}}\rangle}_{Nl_{\max}\ \text{qubits}} \tag{2}$$

with the property that, if the individual code words are all length eigenstates, then U maps the code words to a ZEF string of the $Nl_{\max}$ qubits—that is, one with $|0\rangle$'s after the first $L=l_1+\cdots+l_N$ qubits:

$$|\psi_1^{1\cdots l_1}0^{l_1+1\cdots l_{\max}}\rangle\otimes\cdots\otimes|\psi_N^{1\cdots l_N}0^{l_N+1\cdots l_{\max}}\rangle \rightarrow |\Psi^{1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle. \tag{3}$$

This process is called *condensation*. Since every code word is a superposition of length eigenstates, it suffices to specify how the condensation process functions for such code words.

Note that we have made no assumptions about the details of the condensation process. In the most straightforward case, condensation would be accomplished by concatenation of the code words. The condensed state in Eq. (3) would be of the form

$$|\Psi^{1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle = |\psi_1^{1\cdots l_1}\cdots\psi_N^{L-l_N+1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle. \tag{4}$$

This special type of condensation is called *simple condensation*, and those codes whose code words can be condensed in this way are said to be *simply condensable* codes. Obviously, all simply condensable codes are condensable; but the converse is not true.

The condensability condition is phrased as an ''encoding'' requirement, but the unitary character of the packing process automatically yields a decoding condition—we can unpack a condensed string by applying the U^{-1} transformation.

It is interesting to compare the analogous classical situation. Classical code words in a variable-length code can always be concatenated into a ''packed'' string. Only for uniquely decipherable codes is this packing reversible. In the quantum case, since arbitrary superpositions of code words are also legal code words, the concatenation process itself must be unitary. This automatically implies that it can be reversed.

C. Quantum Kraft-McMillan inequality

Given that the code words carry their own length information and form a condensable code, we next derive a condition on the code word length observable. Fix a value of N and consider all code word strings that have given values of $l_1,l_2,\ldots,l_N$. These states lie in a subspace of dimension $d_{l_1}d_{l_2}\cdots d_{l_N}$, and all of them are mapped by U into something of the form $|\Psi^{1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle$.

Next, imagine strings of code words with different lengths $l_1',l_2',\ldots,l_N'$, but whose lengths sum to the same total length: $L'=L$. The space spanned by these has dimension $d_{l_1'}d_{l_2'}\cdots d_{l_N'}$ and is orthogonal to the previous space. We can consider all such combinations of lengths that sum to the same L. Each of these states maps under U to something of the form $|\Psi^{1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle$, so we obtain (dimension of space containing all code word strings with the same L)$\leqslant$(dimension of space containing all strings $|\Psi^{1\cdots L}0^{L+1\cdots Nl_{\max}}\rangle$),

$$\sum_{l_1+\cdots+l_N=L} d_{l_1}\cdots d_{l_N} \leqslant 2^L.$$

It follows that

$$2^{-L}\sum_{l_1+\cdots+l_N=L} d_{l_1}\cdots d_{l_N} = \sum_{l_1+\cdots+l_N=L} (2^{-l_1}d_{l_1})\cdots(2^{-l_N}d_{l_N}) \leqslant 1.$$

There are at most $Nl_{\max}$ possible values of L. If we sum both sides of this equation over those values, the resulting sum on the left-hand side will include all possible values of $l_1,\ldots,l_N$. Therefore,

$$\sum_{l_1,\ldots,l_{\max}} (2^{-l_1}d_{l_1})\cdots(2^{-l_N}d_{l_N}) = \left(\sum_l 2^{-l}d_l\right)^N \leqslant N l_{\max}.$$

This is of the form $K^N \leqslant N l_{\max}$. If $K>1$, then this inequality must be violated for sufficiently large N. Thus, we conclude that $K \leqslant 1$. But

$$K = \sum_l 2^{-l} d_l = \sum_l 2^{-l} \operatorname{Tr} \pi_l = \operatorname{Tr}\left(\sum_l 2^{-l}\pi_l\right).$$

This gives us our quantum version of the Kraft–McMillan inequality. For any indeterminate-length quantum code that is condensable, the length observable Λ on ZEF code words must satisfy

$$\operatorname{Tr} 2^{-\Lambda} \leqslant 1 \tag{5}$$

(where the trace is taken over the subspace of ZEF code words).

D. Prefix-free codes

An alternate condition that we might impose on our indeterminate-length quantum code is that the code be *prefix free*—informally, that no initial segment of a ZEF code word is itself a code word. In the next Sec. III, we will show that all prefix-free codes are simply condensable. In this section, we will discuss the meaning of the prefix-free condition and show that any condensable code can be transformed into a prefix-free code with the same length characteristics.

Suppose $|\psi_1\rangle$ and $|\psi_2\rangle$ are length eigenstate ZEF code words with lengths l_1 and l_2, respectively; and further suppose that $l_2 > l_1$. These states have the form

$$|\psi_1\rangle = |\psi_1^{1\cdots l_1} 0^{l_1+1\cdots l_{\max}}\rangle,$$

$$|\psi_2\rangle = |\psi_2^{1\cdots l_2} 0^{l_2+1\cdots l_{\max}}\rangle. \tag{6}$$

For the code word $|\psi_1\rangle$, the quantum state of the first l_1 qubits of the register is just the pure state $|\psi_1^{1\cdots l_1}\rangle$. For the code word $|\psi_2\rangle$, the first l_1 qubits may be in a mixed state, described by the density operator

$$\rho_2^{1\cdots l_1} = \operatorname{Tr}_{l_1+1\cdots l_{\max}} |\psi_2\rangle\langle\psi_2| = \operatorname{Tr}_{l_1+1\cdots l_2} |\psi_2^{1\cdots l_2}\rangle\langle\psi_2^{1\cdots l_2}|. \tag{7}$$

We say that our code is *prefix free* if, for all such pairs of code words,

$$\langle\psi_1^{1\cdots l_1}|\rho_2^{1\cdots l_1}|\psi_1^{1\cdots l_1}\rangle = 0. \tag{8}$$

In other words, the first l_1 qubits of a code word of length l_1 have a state that is orthogonal to the (possibly mixed) state of the first l_1 qubits of a code word of length $l_2 > l_1$.

Another way of expressing this condition is to say that a length eigenstate ZEF code word of length l can be distinguished from a code word of greater length by making a measurement only on the first l qubits. Shorter code words can be ''recognized'' from shorter segments. This means that the operator π_l, the projection onto the subspace of length eigenstate ZEF code words of length l, has the form

$$\pi_l = \pi^{1\cdots l} \otimes 1^{l+1\cdots l_{\max}}. \tag{9}$$

Of course, actually to measure the code word length Λ would be disastrous, because such a real measurement would destroy the coherence of superpositions of length eigenstates without possibility of restoration. The condensation process must therefore not include any measurement of length information. On the other hand, the process may include interactions by which, at some intermediate stage, a quantum computer has become entangled with code word length information—provided that, by the end of the computation, this entanglement has been eliminated. In Sec. I E, we discuss this in more detail.

A particularly simple way of generating a prefix-free quantum code is to use a classical prefix-free code as a basis for the ZEF code word subspace. For example, the classical code words 0, 10, 110, and 111 form a prefix-free set. The corresponding quantum code can be specified by giving an orthogonal basis of length eigenstate ZEF code words, as follows:

state	length
$\|000\rangle$	1
$\|100\rangle$	2
$\|110\rangle$	3
$\|111\rangle$	3

The length observable Λ for this code is

$$\Lambda = |000\rangle\langle 000| + 2|100\rangle\langle 100| + 3|110\rangle\langle 110| + 3|111\rangle\langle 111|. \tag{10}$$

Of course, any superposition of these is also a ZEF code word, though not necessarily a code word of definite Λ. It is easy to verify that the code defined in this way satisfies the criterion for a quantum prefix-free code. The procedure illustrated here may be extended in the obvious way to create a quantum prefix-free code from any classical prefix-free code.

Suppose we have an indeterminate-length quantum code that satisfies the quantum Kraft-McMillan inequality. Then, the space of ZEF code words of this code is spanned by a basis of eigenstates of the code's length observable Λ of the code. Let $|\psi_{\mathrm{ZEF};l,i}\rangle$ be the ith such basis vector with length eigenvalue l, and let n_l be the number of basis vectors that have length l. (Thus, for a given l, i ranges from 1 to n_l.) The quantum Kraft-McMillan inequality [Eq. (5)] implies that

$$\sum_l n_l 2^{-l} \leqslant 1. \tag{11}$$

Given values of l and n_l that satisfy Eq. (11), we can construct a classical prefix-free code with n_l distinct code words of length l bits. [In this case, Eq. (11) is just the classical Kraft inequality.] Denote by $C_{l,i}$ the ith code word of length l bits in this prefix-free code. We use this classical

prefix-free code to create a quantum prefix-free code by constructing a basis of length eigenstates, whose elements are $|C_{l,i}0^{l+1\cdots l_{\max}}\rangle$.

Now consider the mapping

$$|\psi_{\mathrm{ZEF};l,i}\rangle \rightarrow |C_{l,i}0^{l+1\cdots l_{\max}}\rangle. \tag{12}$$

This is a mapping from orthogonal basis vectors to orthogonal basis vectors that can be extended linearly to a unitary mapping V on the entire Hilbert space. V takes the original code words to prefix-free code words in a length-preserving way—that is, the length observable Λ' of the prefix-free code is given by $\Lambda' = V\Lambda V^{\dagger}$. In short, any quantum code that satisfies Eq. (5) can be unitarily mapped to a prefix-free quantum code with identical length characteristics.

Are all prefix-free quantum codes condensable? As we shall see in Sec. I E, they are; but in order to show this, we will have to give an explicit algorithm for a quantum computer to condense the code words of a prefix-free quantum code. This algorithm must maintain the coherence of superpositions of code words of different lengths. Before we describe our algorithm, we will first discuss some key characteristics of coherent information processing.

E. Coherence and reversibility

We adopt a high-level model of a quantum computer, which could in principle be implemented by a quantum Turing machine or an array of quantum gates. Our quantum computer contains several registers of qubits, which initially hold ZEF code words from a prefix-free quantum code. The computer also includes a central processing unit that contains various counters and pointers, each of which can take on integer values (or superpositions of these). A system clock keeps track of the number of machine cycles that have passed since the beginning of the computation. (This clock may be treated as an entirely classical system; its function is simply to control the execution of our quantum program.) Finally, the computer contains an output ''tape'' of qubits (initially all in the state $|0\rangle$) on which the condensed string is to be written.

Our job is to write the input code words onto the output tape in a way that preserves the coherence of superpositions of different code words, including superpositions of code words of different lengths. This means that the operation of the computer must be unitary. We can guarantee this unitarity if we satisfy certain conditions:

(1) *Reversibility*. In a classical code, all code words have a determinate length. We can choose an orthogonal basis of length eigenstates to be ''quasiclassical'' input states of our computer. (These states need not be fully classical—for example, the qubits in these code word states may be entangled with each other. However, each code word in our basis has a determinate length.) We require that distinct quasiclassical inputs lead to distinct final states of the computer. This is essentially a requirement that the computation be *reversible* on these quasiclassical inputs [10,11].

(2) *Coherent computation*. The computation includes no measurement or process in which the environment becomes entangled with the computer. As a special case of this, we require that the computation end after exactly the same number of steps for any input code word. If the computation took more steps for longer code words, the halting time of the computation would constitute a measurement of code word length, and would destroy the coherence.

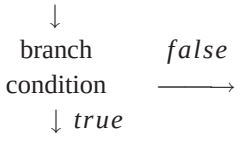
(3) *Localization of coherence in the output*. For any quasiclassical input, at the end of the computation all input registers and internal variables in the central processor have been reset to fixed values that are independent of the input. Only the output tape retains any information about the input. This will guarantee that a superposition of quasiclassical inputs will not lead to entanglement between the output tape and the rest of the computer; the coherence will be localized in the output tape.

A similar set of conditions is outlined in Ref. [6], where it is used to specify quantum algorithms for data compression and for quantum arithmetic coding.

The reversibility requirement ensures that an orthogonal basis of initial states maps to an orthogonal basis of final states. If the computation is coherent, this map extends by linearity to a unitary evolution for the quantum state of the computer. The final requirement guarantees that the quantum information initially in the input registers can be recovered from the condensed output tape alone. We will discuss each of these requirements in turn.

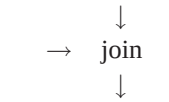
Consider how our quantum computer acts on quasiclassical (length eigenstate) inputs. If we were to map out its algorithm as a flowchart, the requirement of reversibility would impose two sorts of requirements. First, each individual operation on the data must be reversible. Second, the branches and joins in the flowchart must be specified in a reversible way.

A branch can be pictured in this way:

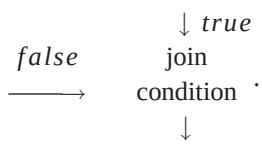


Execution of the program enters from the top, and a logical ''branch condition'' is evaluated. If the branch condition is true, execution proceeds along the downward branch; if false, along the rightward branch. This is plainly reversible, as long as the evaluation of the branch condition is done in a reversible way; there is no ambiguity in the execution of the reversed program.

However, a simple join

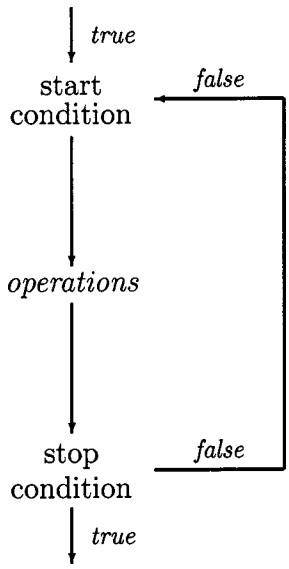


is not reversible, since in the reversed program it is not clear which of the two paths to take. The point is that a join in the flowchart is a reversed branch, and thus must be governed by a logical ''join condition'':



The program is designed so that the join condition is true whenever the execution approaches from above, and false whenever execution approaches from the right.

In our program, we will want to use branches and joins to create ''loops,'' like so:



The ''start condition'' is a logical condition that is only true at the beginning of the execution of the loop and not thereafter; the ''stop condition'' is only true at the end of the execution of the loop and not before.

We can also conveniently represent the reversible loop structure in pseudocode form:

loop enter (*start condition*)

operations

loop exit (*stop condition*).

Both the beginning and the end of the loop are governed by logical conditions.

The requirement that the computation be coherent may at first seem difficult to achieve, since each branch point (or join point) in the algorithm involves the evaluation of a condition—apparently a measurement process. However, these conditions can control the execution of the program without any irreversible loss of coherence.

Let us suppose that the quantum system Q is some portion of our computer, and that we wish to branch our program based on a condition about the state of Q. The condition is represented by a projection Π acting on the Hilbert space $\mathcal{H}$ describing Q. Any initial state $|\psi\rangle$ of Q can be written as

$$|\psi\rangle = \Pi|\psi\rangle + \Pi^{\perp}|\psi\rangle. \tag{13}$$

We could imagine evaluating the condition by making a measurement of the observable represented by Π and $\Pi^{\perp}$. But this would destroy the coherence in this superposition, so a less destructive operation is required.

We join to Q a single qubit (in another part of the quantum computer), and consider the operator U on joint system:

$$U = (|0\rangle\langle 1| + |1\rangle\langle 0|) \otimes \Pi + 1 \otimes \Pi^{\perp}. \tag{14}$$

U is easily verified to be a unitary operator, and thus it could represent some coherent quantum evolution of the joint system. If the qubit is initially set to the state $|0\rangle$ and then U acts, we obtain

$$U|0\rangle \otimes |\psi\rangle = |1\rangle \otimes \Pi|\psi\rangle + |0\rangle \otimes \Pi^{\perp}|\psi\rangle. \tag{15}$$

This is an entangled state of the qubit and Q. If we were to make a measurement of the qubit in the standard basis, we would be effectively measuring the observable Π on Q. That is, the qubit ''contains'' the value of the observable Π. However, the interaction is completely reversible, and in this case may be undone by a further application of U itself.

The qubit can be used as a switch to instruct the computer which branch of the computation to follow. Suppose we wish to specify that, if the qubit is $|0\rangle$, the rest of the computer performs a computation described by the unitary operator V_0, whereas if the qubit is $|1\rangle$ then we wish to do the computation V_1. Then we instruct the entire computer (including the switch qubit) to perform a coherent computation described by the unitary operator

$$V = |0\rangle\langle 0| \otimes V_0 + |1\rangle\langle 1| \otimes V_1. \tag{16}$$

If the overall state of the computer is a superposition of the two switch states, both branches are followed in different branches of the superposition. The computer may become increasingly entangled, but the coherence of its overall state is preserved.

We have shown that any branching condition that can be represented by a projection operator Π can be used to control the execution of the program without any necessary loss of coherence. The cost is entanglement among the parts of the computer.

A join point in the algorithm is simply a time-reversed branch point. Just before the join, the computer is in a state like Eq. (15), in which the qubit is entangled with the system Q. The operator $U^{-1} = U$ acts, and we return the qubit to the state $|0\rangle$ and the system Q to a state like Eq. (13). We have ''disentangled'' Q from the qubit, so the two branches of the computation (controlled by the qubit) have merged.

Our second concern with coherent computation is the synchronization of the computation on different components of the initial superposition. This can be maintained without much difficulty by introducing appropriate ''delay loops'' into the program, so that its execution requires exactly the same number of machine cycles for any input.

We will address our final concern, that the output tape should wind up unentangled with the rest of the computer, by

showing that the final state of the rest of the computer (input registers and central processor) is independent of the input state.

F. Prefix-free codes are simply condensable

We are at last ready to give our algorithm for simply condensing the code words of a prefix-free quantum code. First, we establish our notation and describe the contents of our computer in slightly more detail.

1. Registers

Our computer contains N registers, each consisting of $l_{\max}$ qubits. The ith register is denoted R_i and the kth qubit of this register is called $R_{i,k}$. Initially, each register contains a ZEF code word from a fixed prefix-free quantum code.

2. Tape

There is a tape T containing at least $Nl_{\max}$ qubits, all of which are initially in the state $|0\rangle$. The nth qubit in the tape is called T_n.

3. Counter

There is a counter variable c, which can take on integer values starting with 0 (or, of course, superpositions of these). The initial state of c is $|0\rangle$.

4. Pointers

There are several pointer variables, which like the counter variable take on integer values and have an initial state $|0\rangle$. These variables point to locations in the memory of the computer, but of course they are themselves quantum variables and can take on entangled superpositions of values. There is an overall register pointer r and, for each register, a qubit pointer q_i (for the ith register). The tape also has a pointer variable p.

Once again we emphasize that all of these components, and the operations they perform, can be simulated by a quantum Turing machine or an array of quantum gates. We distinguish among them only to make clear the structure of the condensation algorithm.

The first section of the program copies the contents of the registers to the tapes, moving the pointers in the process:

loop enter $(r=0)$
  $r \leftarrow r+1$
  loop energy $(q_r=0)$
    $q_r \leftarrow q_r+1$
    $p \leftarrow p+1$
    $T_p \leftarrow T_p \oplus R_{r,q_r}$
  loop exit $(R_{r,1}\cdots R_{r,q_r}$ is a code word of length $q_r)$
loop exit $(r=N)$.

(The notation $a \leftarrow a \oplus b$ indicates the ''controlled not'' operation on the qubits, with a as the ''target'' and b as the ''control'' qubit.) Notice that the exit condition for the inner loop (which copies the register qubits one by one onto the tape) is legitimate because the code is prefix free. This means that the question of whether the first q_r qubits of the register form a code word of length q_r can be settled by measuring a projection-type observable on those qubits. (The computer does not make such a measurement, but instead coherently controls its operation as just described.)

We also note that, since the procedure is just to copy the register contents to the output tape, we are doing simple condensation.

At this stage, the various pointer variables are entangled with the code word length information; furthermore, the time at which the computer reaches this stage of the computation is indeterminate. We now resynchronize the program via a delay loop that causes the computer to ''idle'' until a fixed time D (chosen large enough so that the first section of the program has finished even for the longest possible input code words):

loop enter $(c=0)$
  $c \leftarrow c+1$
loop exit (time $=D$).

The second half of the program is the reverse of the first half, except that the register is uncopied, rather than the tape.

loop enter (time $=D+1$)
  $c \leftarrow c-1$
loop exit $(c=0)$
loop enter $(r=N)$
  loop enter $(R_{r,1}\cdots R_{r,q_r}$ is a code word of length $q_r)$
    $R_{r,q_r} \leftarrow T_p \oplus R_{r,q_r}$
    $p \leftarrow p-q$
    $q_r \leftarrow q_r-1$
  loop exit $(q_r=0)$
  $r \leftarrow r-1$
loop exit $(r=0)$.

The program now ends, after exactly $2D$ machine steps. All pointers and counters have been returned to their initial zero values, and the input qubit registers have been reset to $|00\cdots 0\rangle$. Only the qubit tape now contains any nonzero data, in the form of a simply condensed string of N code words. In short, the computer at the end retains no code word-length information at all. Superpositions of code words of different length will thus remain coherent in the condensation process. Since the algorithm works for any given N, the prefix-free quantum code is simply condensable.

We previously proved that every condensable code satisfies the quantum Kraft-McMillan inequality, and then that every quantum code that satisfies the Kraft-McMillan inequality can be unitarily remapped to a prefix-free code. We now learn that prefix-free quantum codes are simply condensable. Since unitary remapping might be part of a general condensation process, we have established that a quantum code is condensable if and only if it satisfies the quantum Kraft-McMillan inequality.

III. QUANTUM DATA COMPRESSION

A. How many qubits?

Classical variable-length codes are used for *data compression*—that is, the representation of classical information in a compact way, using as few resources (bits) as possible. This is done by encoding more probable messages in shorter code words, so that the average code word length is minimized. In this section, we will discuss how—and in what sense—quantum indeterminate-length codes may be used for quantum data compression.

Suppose Alice is sending classical information to Bob using the following classical variable-length code:

message	code word
C_1	0
C_2	10
C_3	110
C_4	111.

If the message C_1 is sent, Bob receives a signal consisting of a single bit (0); but if C_4 is sent, he receives three bits (111). In each case, Bob knows how many bits are being used to send the message. If a long string of messages is being sent, Bob at any stage knows how many complete messages have been received.

Bob learns the length of each code word because he actually learns which code word was sent. The fact that Bob learns the identity of each code word is not a problem in the classical situation; indeed, it is the whole point of classical communication! This contrasts with quantum information transfer. If the signals from Alice, for example, are drawn from a nonorthogonal set of states, Bob will not be able to determine reliably which signal was sent, and any attempt to do so would damage the fidelity of the quantum information.

Suppose that Alice wishes to send quantum information to Bob using the quantum analog of the aforementioned prefix-free code. In other words, the length eigenstate ZEF code words are

state	length
$\|000\rangle$	1
$\|100\rangle$	2
$\|110\rangle$	3
$\|111\rangle$	3

Arbitrary superpositions of these code words are also allowed code words. To maintain the coherence of these superpositions, therefore, Bob must not obtain any information about the length of the code word he receives.

A quantum system actually used for the transmission of information must have at least two degrees of freedom. The first is the ''data'' degree of freedom, which may for instance be a qubit. The second degree of freedom is the ''location'' degree of freedom. This is the physical degree of freedom which determines whether or not Bob has access to the data degree of freedom. The faithful transmission of a qubit in a state $|\psi\rangle$ from the location of Alice a to the location of Bob b would be a process like this:

$$|\psi,a\rangle \rightarrow |\psi,b\rangle. \tag{17}$$

Although we are phrasing our discussion in terms of the transmission of quantum information from one spatial location to another, this analysis would also apply to the storage and retrieval of information in a quantum computer. There, the location degree of freedom might be the reading of a clock; the information stored at time a is to be retrieved at some later time b.

If we have several data qubits, each one will have a location degree of freedom (which may, of course, be correlated with the others). The number of qubits transmitted from Alice to Bob will be the number of location degrees of freedom that have evolved from a to b. For instance, suppose that three data qubits are in a joint state $|\psi^{123}\rangle$, and that Alice sends the first and third qubits to Bob. The final state would be $|\psi^{123},bab\rangle$, in which Bob has received two qubits.

How could Alice send an *indeterminate* number of qubits to Bob—in particular, if Alice is representing her quantum information using the prefix-free quantum code above, how can she arrange to send only the first l qubits of a ZEF code word of length l? The transmission of the length eigenstates is easy to describe:

$$|000,aaa\rangle \rightarrow |000,baa\rangle$$

$$|100,aaa\rangle \rightarrow |100,bba\rangle$$

$$|110,aaa\rangle \rightarrow |110,bbb\rangle$$

$$|111,aaa\rangle \rightarrow |111,bbb\rangle.$$

But imagine that Alice is sending a superposition of code words of different lengths. If this process is unitary, then at the end the data qubits will be entangled with their location degrees of freedom. The coherence of the superposition would no longer be maintained within the data qubits. In order to restore the coherence, Bob would have to interact with the location degrees of freedom of the qubits with which he has indeterminate access. Except for a trivial case—in which Bob simply returns the qubits from location b back to a—he will not be able to do this.

If the transmission process is not unitary, things are even worse. Our conclusion is that it is not possible to send quantum information coherently using an indeterminate number of qubits. If we are to use indeterminate-length quantum codes for quantum data compression, we will have to do so in such a way that a *fixed* number of qubits changes hands from Alice to Bob.

Perfect fidelity would demand that Alice send *all* of the qubits to Bob—enough qubits so that even the longest component of each code word is transmitted in its entirety. But this scheme would allow for no data compression at all.

Our previous discussion of condensability offers some hope. The condensation process took the ''information-bearing'' parts of N ZEF code words (in registers of length $l_{\max}$) and unitarily shifted them as far as possible toward the

beginning of a tape of $Nl_{\max}$ qubits. Although some branches of the overall superposition may extend to the end of the tape, the "typical" branch may be much shorter (followed by $|0\rangle$'s). We therefore might be able to truncate the condensed string of code words after some number L of qubits, where $L \ll Nl_{\max}$, and still maintain an average fidelity approaching unity.

Let us consider a quantum information source that produces an ensemble of signal states of some quantum system. These signal states are unitarily encoded as ZEF code words of some condensable quantum code. For our purposes, therefore, we can simply consider the ensemble of ZEF code words produced by the quantum information source and the unitary encoding. In this ensemble, the code word $|a_{\mathrm{ZEF}}\rangle$ occurs with probability $p(a)$, and the average encoded signal state is described by the density operator

$$\rho = \sum_a p(a) |a_{\mathrm{ZEF}}\rangle\langle a_{\mathrm{ZEF}}|. \tag{18}$$

Our source produces a sequence of independent, identically distributed signals, which are encoded as ZEF code words in separate registers. The average state of N of these registers is $\rho^{\otimes N}$.

The average length $\langle l \rangle$ of the code word ensemble is

$$\langle l \rangle = \mathrm{Tr}\, \rho \Lambda = \sum_a p(a) \langle a_{\mathrm{ZEF}}| \Lambda |a_{\mathrm{ZEF}}\rangle. \tag{19}$$

The average length $\langle l \rangle$ is an ensemble average of quantum expectation values for Λ, but no code word $|a_{\mathrm{ZEF}}\rangle$ need be a length eigenstate.

A condensed string of N code words is a ZEF string of $Nl_{\max}$ qubits, with length observable $\mathbf{\Lambda}$. If U is the unitary operator that maps the N separate ZEF code words to the condensed string, then we can define the overall length observable for the condensed string to be

$$\mathbf{\Lambda} = U(\Lambda_1 + \Lambda_2 + \cdots + \Lambda_N) U^{-1}.$$

The condensed length $\mathbf{\Lambda}$ is just the sum of the individual length observables of the separate, precondensed code words. This observable will have eigenvalues $L = l_1 + \cdots + l_N$ and an average value $\langle L \rangle$. The code words are independent, and so

$$\langle L \rangle = N \langle l \rangle. \tag{20}$$

Since the overall length of the condensed string is defined to be additive, we can apply the "law of large numbers" to some measurement of $\mathbf{\Lambda}$: For any $\epsilon, \delta > 0$, for large enough N it is true that

$$\Pr(|\mathbf{\Lambda} - N\langle l \rangle| > N\delta) < \epsilon. \tag{21}$$

This means that, for large N, the probability is very small that $\mathbf{\Lambda}$ will be found to be much less than (or much greater than) $\langle L \rangle$. Of course, we will not in general make such a measurement, but Eq. (21) is still useful in restricting the typical amplitude of code word string components.

As we shall see, if the ensemble average length of the ZEF code words is $\langle l \rangle$, then we can in the long run maintain fidelity near to 1 by keeping just $\langle l \rangle + \delta$ qubits per signal, where δ can be made as small as desired. Conversely, in a simple condensation process, we must keep at least $\langle l \rangle$ qubits per signal to maintain high fidelity—if we keep only $\langle l \rangle - \delta$ per signal, the average fidelity tends toward zero. We will also find that the ensemble average length of the ZEF code words is related to the von Neumann entropy of the signal ensemble, making this approach an alternate route to the noiseless quantum coding theorem. Finally, we will show that the relative entropy is a measure of the additional resources (qubits) required to represent quantum information using a code that is not optimal.

B. Enough qubits

In this section we will make use of the fact that a condensed string of N ZEF code words is itself in ZEF form—in other words, we can view the condensed string as a ZEF code word in a much longer code. The length observable for this super-code-word will be the sum of the length observables for the N original code words.

Suppose we have a ZEF code word $|\phi\rangle$ in a register of n qubits, and suppose $\ell \leq n$. Define η such that a measurement of the length observable Λ on the code word yields a result larger than ℓ with probability

$$\Pr(\Lambda > \ell) = \eta. \tag{22}$$

In general $|\phi\rangle$ will include components of various lengths. Let Π_ℓ be the projection

$$\Pi_\ell = 1^{1 \cdots \ell} \otimes |0^{\ell+1 \cdots n}\rangle\langle 0^{\ell+1 \cdots n}|. \tag{23}$$

That is, Π_ℓ projects onto the subspace of register states that are $|0\rangle$ in the last $n-\ell$ qubits. We can write our ZEF code word $|\phi\rangle$ as

$$|\phi\rangle = \alpha |\phi_{(\leq \ell)}\rangle + \beta |\phi_{(> \ell)}\rangle, \tag{24}$$

where $\alpha, \beta \geq 0$, and $|\phi_{(\leq \ell)}\rangle$ and $|\phi_{(> \ell)}\rangle$ are normalized states such that

$$\Pi_\ell |\phi_{(\leq \ell)}\rangle = |\phi_{(\leq \ell)}\rangle,$$

$$\Pi_\ell |\phi_{(> \ell)}\rangle = 0.$$

Since all Λ-eigenstate ZEF code words with length no larger than ℓ have $|0\rangle$ in the last $n-\ell$ qubits,

$$1 - \eta = \Pr(\Lambda \leq \ell) \leq \alpha^2. \tag{25}$$

Equality need not hold, however, since some length eigenstate code words with $\Lambda > \ell$ may nevertheless have $|0\rangle$ in the last $n-\ell$ qubits. (This is analogous to the classical situation, in which it is perfectly possible to have one or more 0's at the end of a code word in a variable-length code.)

We now imagine that we truncate the register by discarding the last $n-\ell$ qubits. Only ℓ qubits are stored or transmitted. At the end of the receiver of the process, $n-\ell$ qubits

in the standard state $|0\rangle$ are appended, yielding a mixed final state σ for the register. With what fidelity $F=\langle\phi|\sigma|\phi\rangle$ has the original code word state been maintained by this process?

Direct calculation shows that the mixed state σ is

$$\sigma=\alpha^2|\phi_{(\leqslant\ell)}\rangle\langle\phi_{(\leqslant\ell)}|+\beta^2 w_{(>\ell)}, \tag{26}$$

where $w_{(>\ell)}$ is the state obtained by truncating $|\phi_{(>\ell)}\rangle$ and appending $n-\ell$ qubits in the state $|0\rangle$. Thus

$$F=\langle\phi|\sigma|\phi\rangle=\alpha^2|\langle\phi|\phi_{(\leqslant\ell)}\rangle|^2+\beta^2\langle\phi|w_{(>\ell)}|\phi\rangle\geqslant\alpha^4. \tag{27}$$

Therefore,

$$F\geqslant\alpha^4\geqslant(1-\eta)^2\geqslant 1-2\eta. \tag{28}$$

If the code word length Λ would be found to be no more than ℓ with probability $1-\eta$, then we can keep only ℓ qubits and recover the original state with fidelity $F\geqslant 1-2\eta$.

We can now apply this result and the law of large numbers [Eq. (21)] to a condensed string of code words. If $\epsilon,\delta>0$ and N is sufficiently large, and if we take $\ell=N(\langle l\rangle+\delta)$, then the ensemble average probability that the code word string is longer than ℓ can be made smaller than $\epsilon/2$. We can therefore truncate the string after only $N(\langle l\rangle+\delta)$ qubits and later recover the original string with an average fidelity

$$\langle F\rangle>1-\epsilon. \tag{29}$$

Therefore, if we keep more than $\langle l\rangle$ qubits per input message, in the long run we will be able to retrieve the quantum information with average fidelity approaching unity. The average length $\langle l\rangle$ tells us how many qubits are sufficient for high fidelity.

C. Too few qubits

We now turn to the question of how many qubits are necessary to achieve high fidelity after the condensed string is truncated. For this discussion we will restrict our attention to *simple condensation*, rather than a general condensation process. Since any condensable code can be replaced by a simply condensable code with the same length characteristics, this restriction is not too severe.

The reason for making this restriction is pragmatic. Suppose we have N registers containing code words from a condensable code, with an average length of $\langle l\rangle$. A general condensation procedure might consist of two stages. In the first, the code words in the N separate registers are unitarily remapped to code words from a more efficient code, that is, one with shorter average length $\langle l'\rangle<\langle l\rangle$. In the second stage, this more efficient code is condensed. We have established that only about $N\langle l'\rangle$ qubits will be sufficient to maintain high fidelity. In other words, the original average length $\langle l\rangle$ may tell us nothing about the number of qubits necessary for high fidelity.

Of course, we might not choose to condense the code words in this way, or a more efficient code might not exist. Our strategy will be to separate the question of the efficiency of a code from the question of how many qubits are necessary. First, we will consider the simple condensation of codes that may be inefficient, and then (in the next section) we will discuss limits on the efficiency of codes. In this section, therefore, we describe limits imposed by the structure of our particular (possibly suboptimal) code, and in the next we will indicate how optimal or near-optimal codes may be chosen.

Begin with N ZEF code words of a simply condensable code. The simply condensed string formed from the N code words can be built out of two pieces: (1) the simply condensed qubit string obtained from the first $N-k$ code words, and (2) the simply condensed qubit string obtained from the last k code words. These two pieces are both ZEF and are simply condensed together to form the complete string. Thus, we will base our discussion on the simple condensation of just two ZEF code words.

The first ZEF code word $|\psi\rangle$ lies in a register of m qubits, and the second code word $|\chi\rangle$ lies in a register of n qubits. The simply condensed pair (denoted rather symbolically by $|\psi\chi\rangle$) is a state of a string of $m+n$ qubits. We also consider a state called $|\psi 0\rangle$, which is the first ZEF code word followed by n additional qubits in the state $|0\rangle$.

Let $\ell\leqslant m+n$. The first ZEF code word can be written

$$|\psi\rangle=\alpha|\psi_{(<\ell)}\rangle+\beta|\psi_{(\geqslant\ell)}\rangle, \tag{30}$$

where $\alpha,\beta\geqslant 0$ and $|\psi_{(<\ell)}\rangle$ (or $|\psi_{(\geqslant\ell)}\rangle$) is a normalized superposition of length eigenstates that are shorter than (or at least as long as) ℓ. If we now simply condense this code word with the code word $|\chi\rangle$, we obtain

$$|\psi\chi\rangle=\alpha|\psi_{(<\ell)}\chi\rangle+\beta|\psi_{(\geqslant\ell)}\chi\rangle, \tag{31}$$

with $|\psi_{(<\ell)}\chi\rangle$ and $|\psi_{(\geqslant\ell)}\chi\rangle$ being the simply condensed strings obtained from $|\chi\rangle$ and the two components of $|\psi\rangle$. In a similar way,

$$|\psi 0\rangle=\alpha|\psi_{(<\ell)}0\rangle+\beta|\psi_{(\geqslant\ell)}0\rangle. \tag{32}$$

Now, we imagine truncating the string of $m+n$ qubits, keeping only the first ℓ of them to be stored or transmitted. (We can denote this process by $\mathcal{T}_\ell$.) At the end of the receiver, we do some unspecified quantum operation $\mathcal{E}$ that results in a final state of $m+n$ qubits. We know nothing about $\mathcal{E}$ in general except that it is a trace preserving, completely positive linear map on density operators. The overall process, applied to the two initial states $|\psi\chi\rangle$ and $|\psi 0\rangle$, yields

$$|\psi\chi\rangle\overset{\mathcal{T}_\ell}{\rightarrow}\omega\overset{\mathcal{E}}{\rightarrow}\mathcal{E}(\omega),$$

$$|\psi 0\rangle\overset{\mathcal{T}_\ell}{\rightarrow}\sigma\overset{\mathcal{E}}{\rightarrow}\mathcal{E}(\sigma).$$

At the end of this process, we are interested in the overall fidelity of the truncation–*cum*–recovery process:

$$F=\langle\psi\chi|\mathcal{E}(\omega)|\psi\chi\rangle. \tag{33}$$

We will show that, under suitable conditions, this fidelity must be small.

For general density operators, the fidelity is defined to be

$$F(\rho_1,\rho_2)=\max|\langle 1|2\rangle|^2, \tag{34}$$

where the maximum is taken over all purifications $|1\rangle$ of ρ_1 and $|2\rangle$ of ρ_2. (Equivalently, we can fix one of the purifications $|1\rangle$ and maximize over the other purification $|2\rangle$.) The fidelity has the property that it is never decreased by any quantum operation, so that

$$F(\mathcal{E}(\rho_1),\mathcal{E}(\rho_2))\geqslant F(\rho_1,\rho_2) \tag{35}$$

for any trace preserving, completely positive linear map $\mathcal{E}$.

A useful result (shown in Ref. [13]) relates the fidelities among three states ρ_1, ρ_2, and ρ_3. Let $F_{12}=F(\rho_1,\rho_2)$, etc. Then

$$\sqrt{F_{13}}\leqslant\sqrt{F_{23}}+\sqrt{2(1-\sqrt{F_{12}})}. \tag{36}$$

This implies that, if F_{12} is nearly equal to one and F_{23} is close to zero, F_{13} is also close to zero. Recalling that $0\leqslant F\leqslant 1$ for all fidelities, we note that $1-\sqrt{F}\leqslant 1-F$, and thus

$$\begin{aligned}F_{13}&\leqslant F_{23}+2(1-F_{12})+2\sqrt{2F_{23}(1-F_{12})}\\&\leqslant F_{23}+2(1-F_{12})+2\sqrt{2(1-F_{12})}\\&\leqslant F_{23}+2\sqrt{1-F_{12}}+2\sqrt{2}\sqrt{1-F_{12}},\end{aligned}$$

$$F_{13}\leqslant F_{23}+5\sqrt{1-F_{12}}. \tag{37}$$

Since this inequality is linear in both F_{13} and F_{23}, it will be convenient for situations in which we wish to average over an ensemble of ρ_3 states.

We apply Eq. (37) to our situation as follows. The state $\rho_1=|\psi\chi\rangle\langle\psi\chi|$, the original simply condensed string, and the state $\rho_3=\mathcal{E}(\omega)$, the final state of the simply condensed string after the truncation $\mathcal{T}_\ell$ and the recovery operation $\mathcal{E}$. Playing the role of ρ_2 is the state $\mathcal{E}(\sigma)$, the final state obtained by using $|\psi 0\rangle$ as our input. Since the quantum operation $\mathcal{E}$ can never decrease the fidelity between states, $F[\mathcal{E}(\omega),\mathcal{E}(\sigma)]\geqslant F(\omega,\sigma)$. Therefore,

$$F=\langle\psi\chi|\mathcal{E}(\omega)|\psi\chi\rangle\leqslant\langle\psi\chi|\mathcal{E}(\sigma)|\psi\chi\rangle+5\sqrt{1-F(\omega,\sigma)}. \tag{38}$$

The initial states $|\psi\chi\rangle$ and $|\psi 0\rangle$ are purifications of ω and σ, respectively. The fidelity $F(\omega,\sigma)$ is thus

$$F(\omega,\sigma)=\max_{|\phi_\sigma\rangle}|\langle\psi\chi|\phi_\sigma\rangle|^2, \tag{39}$$

where the maximum is taken over all purifications $|\phi_\sigma\rangle$ of σ. Now, all of the purifications of σ are related to one another by unitary operators that act only on the adjoined system, so that

$$F(\omega,\sigma)=\max_U|\langle\psi\chi|(1^{1\cdots\ell}\otimes U^{\ell+1\cdots m+n})|\psi 0\rangle|^2, \tag{40}$$

with the maximum taken over all unitary operators acting on the last $m+n-\ell$ qubits.

We write $|\psi\chi\rangle=\alpha|\psi_{(<\ell)}\chi\rangle+\beta|\psi_{(\geqslant\ell)}\chi\rangle$ and $|\psi 0\rangle=\alpha|\psi_{(<\ell)}0\rangle+\beta|\psi_{(\geqslant\ell)}0\rangle$, as before, and note that, since $|\psi_{(\geqslant\ell)}\rangle$ only contains components of $|\psi\rangle$ that are at least as long as ℓ,

$$\begin{aligned}&\mathrm{Tr}_{\ell+1\cdots m+n}|\psi_{(\geqslant\ell)}\chi\rangle\langle\psi_{(\geqslant\ell)}\chi|\\&\quad=\mathrm{Tr}_{\ell+1\cdots m+n}|\psi_{(\geqslant\ell)}0\rangle\langle\psi_{(\geqslant\ell)}0|.\end{aligned} \tag{41}$$

In this component, the second code word, whose "starting address" in the simply condensed string is entangled with the length of the first code word, lies entirely in the discarded tail of the qubit string. Therefore, there exists a unitary $V^{\ell+1\cdots m+n}$ such that

$$|\psi_{(\geqslant\ell)}\chi\rangle=(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi_{(\geqslant\ell)}0\rangle. \tag{42}$$

Clearly,

$$F(\omega,\sigma)\geqslant|\langle\psi\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi 0\rangle|^2, \tag{43}$$

$$\begin{aligned}&\langle\psi\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi 0\rangle\\&\quad=\alpha^2\langle\psi_{(<\ell)}\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi_{(<\ell)}0\rangle\\&\qquad+\alpha\beta\langle\psi_{(<\ell)}\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi_{(\geqslant\ell)}0\rangle\\&\qquad+\beta\alpha\langle\psi_{(\geqslant\ell)}\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi_{(<\ell)}0\rangle+\beta^2,\\&|\langle\psi\chi|(1^{1\cdots\ell}\otimes V^{\ell+1\cdots m+n})|\psi 0\rangle|\geqslant\beta^2-\alpha^2-2\alpha\beta\\&\quad=1-2\alpha-2\alpha^2\geqslant 1-4\alpha.\end{aligned}$$

Therefore,

$$F(\omega,\sigma)\geqslant(1-4\alpha)^2\geqslant 1-8\alpha. \tag{44}$$

Our overall fidelity must satisfy

$$F\leqslant\langle\psi\chi|\mathcal{E}(\sigma)|\psi\chi\rangle+5\sqrt{8\alpha}\leqslant\langle\psi\chi|\mathcal{E}(\sigma)|\psi\chi\rangle+15\sqrt{\alpha}. \tag{45}$$

Neither the operator $\mathcal{E}(\sigma)$ nor the parameter α depends on the second code word $|\chi\rangle$. We now imagine that the second code word is drawn from an ensemble—that is, that the code word $|\chi\rangle$ occurs with probability $P(\chi)$, so that the ensemble has an average density operator

$$W=\sum_\chi P(\chi)|\chi\rangle\langle\chi|. \tag{46}$$

The average fidelity after truncation $\mathcal{T}_\ell$ and recovery $\mathcal{E}$ will therefore be

$$\overline{F}\leqslant\mathrm{Tr}\,W\mathcal{E}(\sigma)+15\sqrt{\alpha}. \tag{47}$$

Since $\mathcal{E}(\sigma)$ is a positive operator of unit trace, we obtain

$$\overline{F}\leqslant\|W\|+15\sqrt{\alpha}, \tag{48}$$

where $\|W\|$ is the operator norm of W, which (since W is positive) is just the largest eigenvalue of W.

After all of this, we are in a position to apply the law of large numbers [Eq. (21)] again. We will be choosing two large integers, N and k. Our first code word $|\psi\rangle$ in the preceding analysis will be a simply condensed string of $N-k$ code words, and the second code word $|\chi\rangle$ will be a simply condensed string of the remaining k code words. We assume that the code words themselves are drawn from an ensemble with an average state ρ having more than one nonzero eigenvalue—in other words, the ensemble involves more than one code word state.

Let $\epsilon, \delta > 0$. If $\lambda < 1$ is the largest eigenvalue of ρ, then the largest eigenvalue of $\rho^{\otimes k}$ is λ^k. Choose k so that $\lambda^k < \epsilon/2$. Since the last k code words are unitarily condensed into a string with average state W, $\|W\| = \|\rho^{\otimes k}\| < \epsilon/2$.

Now we consider the simply condensed string of the first $N-k$ code words, which we have denoted by $|\psi\rangle$. The length observable for this string is Λ_{N-k}. Given a value of N, we define $\ell = N(\langle l\rangle - \delta)$. We will restrict our attention to values of N large enough so that

$$\ell \leqslant (N-k)\left(\langle l\rangle - \frac{\delta}{2}\right). \tag{49}$$

Applying the law of large numbers, we can now specify N large enough so that $\Pr(\Lambda_{N-k} < \ell) = \alpha^2$ is as small as we like. In particular, we can guarantee that $15\sqrt{\alpha} < \epsilon/2$. Thus,

$$\overline{F} \leqslant \|W\| + 15\sqrt{\alpha} < \epsilon. \tag{50}$$

Therefore, if we keep fewer than $\langle l\rangle$ qubits per input message and use simple condensation, in the long run the fidelity of the retrieved quantum information must approach zero. The average length $\langle l\rangle$ tells us how many qubits are necessary for high fidelity using simple condensation.

D. Entropy and average length

The preceding results provide an interpretation for the average length $\langle l\rangle$ of an indeterminate-length quantum code: $\langle l\rangle$ is just a measure of the resources (qubits) that are both necessary and sufficient to maintain high fidelity of the quantum information, in the situations just described. We now inquire how short $\langle l\rangle$ can be for a given quantum information source. In other words, we will now explore how efficient an indeterminate-length quantum code may be.

Recall the quantum Kraft-McMillan inequality [Eq. (5)]. Any condensable quantum code must have a length observable Λ on ZEF code words that satisfies

$$\mathrm{Tr}\, 2^{-\Lambda} = K \leqslant 1,$$

where the trace is restricted to the ZEF subspace. We can construct a density operator ω on the ZEF subspace by letting

$$\omega = \frac{1}{K} 2^{-\Lambda}. \tag{51}$$

The operator ω, although a positive operator of unit trace, is generally not the same as the ensemble average density operator ρ of the code words produced by the information source.

The average code word length $\langle l\rangle$ is

$$\langle l\rangle = \mathrm{Tr}\,\rho\Lambda = -\mathrm{Tr}\,\rho \log(2^{-\Lambda}) = -\mathrm{Tr}\,\rho \log \omega - \log K.$$

Therefore,

$$\langle l\rangle = S(\rho) + \mathcal{D}(\rho||\omega) - \log K, \tag{52}$$

where $S(\rho)$ is the von Neumann entropy of the density operator ρ

$$S(\rho) = -\mathrm{Tr}\,\rho \log \rho \tag{53}$$

and $\mathcal{D}(\rho||\omega)$ is the quantum relative entropy

$$\mathcal{D}(\rho||\omega) = \mathrm{Tr}\,\rho \log \rho - \mathrm{Tr}\,\rho \log \omega. \tag{54}$$

(We use base-2 logarithms.) The relative entropy has a number of useful properties. For example, it is positive definite, so that $\mathcal{D}(\rho||\omega) > 0$ if and only if $\rho \neq \omega$.

Since $\log K \leqslant 0$,

$$\langle l\rangle \geqslant S(\rho). \tag{55}$$

The average code word length must always be at least as great as the von Neumann entropy of the signal ensemble from the information source.

We can approach this bound by a suitable code. The eigenvalues λ_k of ρ form a probability distribution $\boldsymbol{\lambda}$, and the von Neumann entropy is simply the Shannon entropy of the eigenvalues:

$$S(\rho) = H(\boldsymbol{\lambda}) = -\sum_k \lambda_k \log \lambda_k. \tag{56}$$

The probability distribution $\boldsymbol{\lambda}$ can be used to define a Shannon-Fano code, which is a classical prefix-free binary code whose code words have integer lengths $l_k = \lceil \log \lambda_k \rceil$, so that

$$l_k < \log \lambda_k + 1. \tag{57}$$

This means that the average length of the Shannon-Fano code words satisfies

$$\langle l\rangle = \sum_k \lambda_k l_k < H(\vec{\lambda}) + 1. \tag{58}$$

The classical Shannon-Fano code can be used to define a corresponding prefix-free indeterminate-length quantum code, according to the procedure in Eq. (12). (Such a code was also described by Chuang and Modha in Ref. [6].) Eigenstates of ρ are length eigenstate ZEF code words, and the average code word length satisfies

$$\langle l\rangle < S(\rho) + 1. \tag{59}$$

Asymptotically, this code will achieve high fidelity using about $S(\rho)+1$ qubits per signal.

An alternate scheme is based on Huffman codes, which are classical prefix free codes that actually minimize average code word length $\langle l\rangle$. Equations 58 and 59 are also satisfied for Huffman codes and their quantum versions.

We can do even better if we create our ZEF code words from *blocks* of outputs of the quantum information source. This amounts to considering a new source that produces blocks of n elementary signals, with an ensemble average block state $\rho^{\otimes n}$ having an entropy of $nS(\rho)$. A quantum Shannon-Fano or Huffman code designed for this block source would have an average length of no more than $nS(\rho)+1$, so that we will use only $S(\rho)+1/n$ qubits per elementary signal. Thus, by coding long blocks of signals, we can achieve $\overline{F}\to 1$ with about $S(\rho)$ qubits per elementary signal.

It can be seen that the theory of indeterminate-length quantum codes provides an alternate route to the quantum noiseless coding theorem [6]. The von Neumann entropy $S(\rho)$ measures the physical resources necessary to represent quantum information faithfully.

We now ask: Under what circumstances can we achieve the entropic bound to the code word length exactly, without resorting to block coding? In other words, for what codes and code word ensembles can we have

$$\langle l\rangle = S(\rho)? \tag{60}$$

A code for which this equality holds may be called ''length optimal.'' The answer can be seen from Eq. (52):

$$\langle l\rangle = S(\rho)+\mathcal{D}(\rho||\omega)-\log K.$$

Both $\mathcal{D}(\rho||\omega)$ and $-\log K$ are non-negative, so they must both equal zero for a length-optimal code. In other words,

$$K=\mathrm{Tr}\,2^{-\Lambda}=1 \tag{61}$$

and

$$\rho=\omega=2^{-\Lambda}. \tag{62}$$

A length-optimal code must saturate the quantum Kraft inequality [Eq. (5)], and the code word ensemble must equal the density operator ω constructed from the length observable Λ. Two consequences follow:

(a) Whenever the signal ensemble ρ has only eigenvalues of the form 2^{-m} for integer values of m, we can find a condensable quantum code (with length eigenvalues m) that is length optimal. If ρ has eigenvalues that are not of this form, then no length-optimal code exists.

(b) Some quantum codes saturate the quantum Kraft inequality—for example, those based on classical Huffman codes. These codes will be length-optimal for a code word ensemble with density operator

$$\rho=2^{-\Lambda}. \tag{63}$$

That is, every quantum code that saturates the quantum Kraft inequality is length-optimal for some code word ensemble. If a quantum code does not saturate the quantum Kraft inequality, it is not length-optimal for any code word ensemble.

Suppose we have a code that is length-optimal for some density operator ω; but instead, we use the code for an ensemble of code words described by the density operator ρ. Then the average code word length will be

$$\langle l\rangle = S(\rho)+\mathcal{D}(\rho||\omega). \tag{64}$$

We know that, using block coding, we can asymptotically use as few as $S(\rho)$ qubits to faithfully represent the quantum information produced by the source of ρ. We also know that $\langle l\rangle$ is the minimum number of qubits we need to retain per code word to achieve high fidelity in a simply condensed string of many code words. Thus, the relative entropy $\mathcal{D}(\rho||\omega)$ tells us what additional resources (in qubits) are necessary to faithfully represent the quantum information from the ρ source, if we use a code that is length optimal for a different source (the ''ω source'').

E. Remarks

In the quantum Huffman code of Braunstein *et al.*, code word length information and the code words themselves are stored separately, in entangled strings of qubits. This means that the average number of qubits used to store the quantum information from a given source is increased by an amount logarithmic in the code word length [7]. However, as we have seen, this separate accounting for code word length information is unnecessary. The code words of a quantum indeterminate-length code *carry their own length information*.

This requirement is the basis for Eq. (5), the quantum Kraft-McMillan inequality. We have shown that Eq. (5) is a necessary and sufficient condition for condensability, and further, that any code satisfying Eq. (5) can be unitarily mapped to a prefix-free quantum code with the same length characteristics. Prefix-free codes are themselves simply condensable, and obey the quantum Kraft-McMillan inequality.

Classical prefix-free codes are also called ''instantaneous codes,'' since the receiver of a string of code words can identify an individual code word from the string immediately, before the remainder of the string is received [3]. But this terminology is inapplicable to the quantum case. Suppose we have a simply condensed string of code words from a prefix-free quantum code. The first code word is generally not a length eigenstate, and the length of this code word is entangled with the locations in the qubit string of all subsequent code words. The phase relationship between the different-length components of the first code word is a global property of the state of the entire string. Therefore, in order to coherently recover even the first code word, we will need the entire string (or a sufficiently long initial segment to achieve high overall fidelity). Even prefix-free quantum codes are not ''instantaneous;'' the entire transmission must be completed before any part of it can be ''read.''

The classical Kraft-McMillan inequality [Eq. (11)] arises whenever a set of binary strings satisfies the prefix-free con-

dition. For example, it governs the set of lengths of distinct programs for a classical Turing machine. The Kraft-McMillan inequality therefore plays a central role in the *algorithmic* information theory, in which the information content of a binary string s is defined to be length of the shortest halting program that produces s as its output [12–14]. We may hope that the quantum version of the Kraft-McMillan inequality will serve as a starting point for the development of a quantum algorithmic information theory.

ACKNOWLEDGMENTS

The authors are happy to acknowledge our indebtedness to many colleagues with whom we have discussed this work, including C. M. Caves, S. Braunstein, C. A. Fuchs, W. K. Wootters, T. M. Cover, and I. L. Chuang. One of the authors (B.S.) is grateful for the support of the Rosenbaum Foundation at the Isaac Newton Institute for Mathematical Sciences in the Summer of 1999.

[1] C. H. Bennett, Phys. Today **48**, 24 (1995); C. H. Bennett and D. P. DiVincenzo, Nature (London) **404**, 247 (2000).

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

[3] B. Schumacher, Phys. Rev. A **51**, 2738 (1995); R. Jozsa and B. Schumacher, J. Mod. Opt. **41**, 2343 (1994).

[4] P. W. Shor, Phys. Rev. A **52**, R2493 (1995); A. R. Calderbank and P. W. Shor, *ibid.* **54**, 1098 (1996); A. Steane, Phys. Rev. Lett. **77**, 793 (1996); R. Laflamme, C. Miquel, J. P. Paz, and W. H. Zurek, *ibid.* **77**, 198 (1996).

[5] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Phys. Rev. A **54**, 3824 (1996); H. Barnum, M. A. Nielsen, and B. Schumacher, *ibid.* **57**, 4153 (1998).

[6] B. Schumacher, in Proceedings of the Sante Fe Institute Workshop on Complexity, Entropy, and the Physics of Information (1994) (unpublished).

[7] S. L. Braunstein, C. A. Fuchs, D. Gottesman, and H. -K. Lo, in Proceedings of the 1998 IEEE International Symposium on Information, MIT, Cambridge, MA (unpublished), p. 353; http://xxx.lanl.gov/abs/quant-ph/9805080.

[8] I. L. Chuang and D. S. Modha, IEEE Trans. Inf. Theory **46** (3), 1104 (2000).

[9] K. L Boström, http://xxx.lanl.gov/abs/quant-ph/0009052; http://xxx.lanl.gov/abs/quant-ph/0009073.

[10] C. H. Bennett, IBM J. Res. Dev. **17**, 525 (1973).

[11] T. Toffoli, in *Automata, Languages, and Programming*, edited by W. de Bakker and J. van Leeuwen (Springer, New York, 1980).

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

[13] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, Phys. Rev. A **54**, 4707 (1996).

[14] Ming Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. (Springer, Berlin, 1997).