leaf has been decreased by 1, this new tree is cheaper than the old one, contradicting optimality.

In what follows refer to Figs. 9 and 10 for illustration as we do a case-by-case analysis.

(Case I) If u had a right child but no left one we could simply add its left child to get a new tree with the same cost but fewer bad nodes, contradicting the definition of T. Thus u must have a left child v but no right child. There are two cases.

(Case II) If v is the root of some tree T' then we could move T' to be rooted at the right child of u and leave v a leaf. The new resulting tree has the same cost but fewer bad nodes, again leading to a contradiction.

Otherwise, v is itself a leaf. Let x be the parent of u.

(Case III) If u is a left child of x then we simply remove v, leaving u as a left leaf. The cost of the resulting tree is the same as before but it has one fewer bad node. Again a contradiction.

(Case IV) Otherwise, u is the right child of x and removing u could add a new right child to the tree, possibly even raising its cost. Therefore, in this case we remove *both* u and v. Since x was not bad before (because it is higher than u) removing u does not add a new right leaf to the tree so the cost of the resulting tree remains the same. Since x has now become bad the new tree still has B bad nodes but it has fewer total nodes than T, again causing a contradiction.

We have just seen that there exists some optimal full tree T. We now prove that T is feasible. See Fig. 11 for illustration.

Suppose T is not feasible. Then there exists some right internal node $v \in T$ and left leaf $u \in T$ such that $\mathrm{depth}\,(v) = \mathrm{depth}\,(u)$. Let S be the subtree rooted at v, y the deepest right node $y \in S$, and x the left sibling of y (x and y must exist because T is full). Also suppose that probability p_i is assigned to y. Now detach S from v and attach it to u, erase y and assign p_i to node v. Denote the new tree thus created by T'. Since the only probability whose assigned right leaf has changed is p_i we find that

$$\mathrm{Cost}\,(T') = \mathrm{Cost}\,(T) + (\mathrm{depth}\,(v) - \mathrm{depth}\,(y))p_i.$$

But $\mathrm{depth}\,(v) < \mathrm{depth}\,(y)$ so $\mathrm{Cost}\,(T') < \mathrm{Cost}\,(T)$ contradicting optimality of T. Thus T must be feasible. □

REFERENCES

[1] T. Berger and R. W. Yeung, "Optimum "1"-ended binary prefix codes," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1435–1441, Nov. 1990.

[2] C. Szelok, "Variations of prefix free codes," M.Phil. thesis, Dept. Comput. Sci., Hong Kong Univ. Sci. Technol., Dec. 1997.

[3] R. M. Capocelli, A. D. Santis, L. Gargano, and U. Vaccaro, "On the construction of statistically synchronizable codes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 407–414, Mar. 1992.

[4] R. M. Capocelli, A. D. Santis, and G. Persiano, "Binary prefix codes ending in a "1"," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1296–1302, July 1994.

[5] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. Boston, MA: MIT Press, 1993.

[6] S. Even, *Graph Algorithms*. Rockville, MD: Comput. Sci. Press, 1979.

[7] M. Golin and G. Rote, "A dynamic programming algorithm for constructing optimal prefix-free codes for unequal letter costs," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1770–1781, Sept. 1998.

[8] K. Mehlhorn, *Data Structures and Algorithms 2: Graph Algorithms and NP-Completeness*. Berlin, Germany: Springer-Verlag, 1984.

A Quantum Analog of Huffman Coding

Samuel L. Braunstein, Christopher A. Fuchs, Daniel Gottesman, and Hoi-Kwong Lo

***Abstract*—We analyze a generalization of Huffman coding to the quantum case. In particular, we notice various difficulties in using instantaneous codes for quantum communication. Nevertheless, for the *storage* of quantum information, we have succeeded in constructing a Huffman-coding-inspired quantum scheme. The number of computational steps in the encoding and decoding processes of N quantum signals can be made to be of polylogarithmic depth by a massively parallel implementation of a quantum gate array. This is to be compared with the $O(N^3)$ computational steps required in the sequential implementation by Cleve and DiVincenzo of the well-known quantum noiseless block-coding scheme of Schumacher. We also show that $O(N^2(\log N)^\alpha)$ sequential computational steps are needed for the *communication* of quantum information using another Huffman-coding-inspired scheme where the sender must disentangle her encoding device before the receiver can perform any measurements on his signals.**

***Index Terms*—Data compression, Huffman coding, instantaneous codes, quantum coding, quantum information, variable-length codes.**

I. INTRODUCTION

There has been much recent interest in the subject of quantum information processing. Quantum information is a natural generalization of classical information. It is based on quantum mechanics, a well-tested scientific theory in real experiments. This correspondence concerns quantum information.

The goal of this correspondence is to find a quantum source coding scheme analogous to Huffman coding in the classical source coding theory [3]. Let us recapitulate the result of classical theory. Consider the simple example of a memoryless source that emits a sequence of independent and identically distributed signals each of which is chosen from a list $w_1, w_2, \cdots, w_n$ with probabilities $p_1, p_2, \cdots, p_n$. The task of source coding is to store such signals with a minimal amount of resources. In classical information theory, resources are measured in bits. A standard coding scheme to use is the optimally efficient Huffman coding algorithm, which is a well-known lossless coding scheme for data compression.

Apart from being highly efficient, it has the advantage of being instantaneous, i.e., unlike block coding schemes, the encoding and decoding of each signal can be done immediately. Note also that codewords of variable lengths are used to achieve efficiency. As we will see

Manuscript received May 7, 1999; revised January 11, 2000. This work was supported in part by EPSRC under Grants GR/L91344 and GR/L80676, by the Lee A. DuBridge Fellowship, by DARPA under Grant DAAH04-96-1-0386 through the Quantum Information and Computing (QUIC) Institute administered by ARO, and by the U.S. Department of Energy under Grant DE-FG03-92-ER40701. The work of H.-K. Lo was performed while the author was with Hewlett-Packard Laboratories, Bristol, U.K.

S. L. Braunstein is at SEECS, University of Wales, Bangor LL57 1UT, U.K., and at Hewlett-Packard Labs, Filton Road, Stoke Gifford, Bristol BS34 8QZ, U.K. (e-mail: schmuel@sees.bangor.ac.uk).

C. A. Fuchs is at the Norman Bridge Laboratory of Physics 12-33, California Institute of Technology, Pasadena, CA 91125 USA.

D. Gottesman was with the California Institute of Technology, Pasadena and with the Los Alamos National Laboratory. He is now with Microsoft Research, Microsoft Corporation, Redmond, WA 98052 USA.

H.-K. Lo is with MagiQ Technologies, Inc., New York, NY 10001 USA (e-mail: hkl@magiqtech.com).

Communicated by A. M. Barg, Associate Editor for Coding Theory.

Publisher Item Identifier S 0018-9448(00)04288-7.

below, these two features—instantaneousness and variable length—of Huffman coding are difficult to generalize to the quantum case.

Now let us consider quantum information. In the *quantum* case, we are given a quantum source which emits a time sequence of independent and identically distributed pure-state quantum signals each of which is chosen from $|u_1\rangle$, $|u_2\rangle$, $\cdots |u_m\rangle$ with probabilities $q_1, q_2 \cdots, q_m$, respectively. Notice that $|u_i\rangle$'s are normalized (i.e., unit vectors) but not necessarily orthogonal to each other. Classical coding theory can be regarded as a special case when the signals $|u_i\rangle$ are orthogonal. The goal of quantum source coding is to minimize the number of dimensions of the Hilbert space needed for almost lossless encoding of quantum signals, while maintaining a high fidelity between input and output. For a pure input state $|u_i\rangle$, the fidelity of the output density matrix ρ_i is defined as the probability for it to pass a yes/no test of being the state $|u_i\rangle$. Mathematically, it is given by $\langle u_i|\rho_i|u_i\rangle$ [4]. In particular, we will be concerned with the average fidelity $F = \sum_i q_i\langle u_i|\rho_i|u_i\rangle$ It is convenient to measure the dimensionality of a Hilbert space in terms of the number of qubits (i.e., quantum bits) composing it; that is, the base-2 logarithm of the dimension.

Though there has been some preliminary work on quantum Huffman coding [9], the most well-known quantum source coding scheme is a block coding scheme [10], [5]. The converse of this coding theorem was proven rigorously in [1]. In block coding, if the signals are drawn from an ensemble with density matrix $\rho = \sum q_j\,|u_j\rangle\langle u_j|$, Schumacher coding, which is almost lossless, compresses N signals into $NS(\rho)$ qubits, where $S(\rho) = -\mathrm{tr}\,\rho\,\log\,\rho$ is the von Neumann entropy. To encode N signals *sequentially*, it requires $O(N^3)$ computational steps [2]. The encoding and decoding processes are far from instantaneous. Moreover, the lengths of all the codewords are the same.

II. Difficulties in a Quantum Generalization

A notable feature of quantum information is that measurement of it generally leads to disturbance. While measurement is a passive procedure in classical information theory, it is an integral part of the formalism of quantum mechanics and is an active process. Therefore, the big challenge in quantum coding is: How to encode and decode without disturbing the signals too much by the measurements involved? To illustrate the difficulties involved, we shall first attempt a naive generalization of Huffman coding to the quantum case. Consider the density matrix for each signal $\rho = \sum q_j\,|u_j\rangle\langle u_j|$ and diagonalize it into

$$\rho = \sum_i p_i|\phi_i\rangle\langle\phi_i| \tag{1}$$

where $|\phi_i\rangle$ is an eigenstate and the eigenvalues p_i's are arranged in decreasing order. Huffman coding of a corresponding classical source with the same probability distribution p_i's allows one to construct a one-to-one correspondence between Huffman codewords h_i and the eigenstates $|\phi_i\rangle$. Any input quantum state $|u_j\rangle$ may now be written as a sum over the complete set $|\phi_i\rangle$. Remarkably, this means that, for such a naive generalization of Huffman coding, the length of each signal is a quantum-mechanical variable with its value in a superposition of the length eigenstates. It is not clear what this really means nor how to deal with such an object. If one performs a measurement on the length variable, the statement that measurements lead to disturbance means that irreversible changes to the N signals will be introduced which disastrously reduce the fidelity.

Therefore, to encode the signals faithfully, the sender and the receiver are forbidden to measure the length of each signal. We emphasize that this difficulty—that the sender is ignorant of the length of the signals to be sent—is, in fact, very general. It appears in any distributed scheme of quantum computation. It is also highly analogous to the synchronization problem in the execution of subroutines in a quantum computer: A quantum computer program runs various computational paths simultaneously. Different computational paths may take different numbers of computational steps. A quantum computer is, therefore, generally unsure whether a subroutine has been completed or not. We do not have a satisfactory resolution to those subtle issues in the general case. Of course, the sender can always avoid this problem by adding redundancies (i.e., adding enough zeros to the codewords to make them a fixed length). However, such a prescription is highly inefficient and is self-defeating for our purpose of efficient quantum coding. For this reason, we reject such a prescription in our current discussion.

In the hope of saving resources, the natural next step to try is to stack the signals in line in a single tape during the transmission. To greatly simplify our discussion we shall suppose that the read/write head of the machine is quantum-mechanical with its location given by an internal state of the machine (this head location could be thought of as being specified on a separate tape). But then the second problem arises. Assuming a fixed speed of transmission, the receiver can never be sure when a particular signal, say the seventh signal, arrives. This is because the *total* length of the signals up to that point (from the first to seventh signals) is a quantum-mechanical variable (i.e., it is in a superposition of many possible values). Therefore, Bob generally has a hard time in deciding when would be the correct instant to decode the seventh signal in an instantaneous quantum code.

Let us suppose that the above problem can be solved. For example, Bob may wait "long enough" before performing any measurements. We argue that there remains a third difficulty which is fatal for *instantaneous* quantum codes—that the head location of the encoder is *entangled* with the total length of the signals. If the decoder consumes the quantum signal (i.e., performs measurements on the signals) before the encoding is completed, the record of the total length of the signals in the encoder head will destroy quantum coherence. This decoherence effect is physically the same as a "which path" measurement that destroys the interference pattern in a double-slit experiment. One can also understand this effect simply by considering an example of N copies of a state $a|0\rangle + b|1\rangle$. It is easy to show that if the encoder couples an encoder head to the system and keeps a record of the total number of zeros, the state of each signal will become impure. Consequently, the fidelity between the input and the output is rather poor.

III. Storage of Quantum Signals

Nevertheless, we will show here that Huffman-coding-inspired quantum schemes do exist for both storage and communication of quantum information. In this section we consider the problem of storage. Notice that the above difficulties are due to the requirement of instantaneousness. This leads in a natural way to the question of *storage* of quantum information, where there is no need for instantaneous decoding in the first place. In this case, the decoding does not start until the whole encoding process is done. This immediately gets rid of the second (namely, when to decode) and third (namely, the record in the encoder head) problems mentioned in the last section. However, the first problem reappears in a new incarnation: The *total* length of say N signals is unknown and the encoder is not sure about the number of qubits that he should use. A solution to this problem is to use essentially the law of large numbers. If N is large, then asymptotically the length variable of the N signals has a probability *amplitude* concentrated in the subspace of values between $N(\overline{L} - \delta)$ and $N(\overline{L} + \delta)$ for any $\delta > 0$ [10], [5], [1]. Here $\overline{L}$ is the weighted average length of a Huffman codeword. One can, therefore, truncate

the signal tape into one with a *fixed* length say $N(\overline{L}+\delta)$ ("0's" can be padded to the end of the tape to make up the number if necessary.). Of course, the whole tape is not of variable length anymore. Nonetheless, we will now demonstrate that this tape can be a useful component of a new coding scheme—which we shall call quantum Huffman coding—that shares some of the advantages of Huffman coding over block coding. In particular, assuming that quantum gates can be applied in *parallel*, the encoding and decoding of quantum Huffman coding can be done efficiently. While a sequential implementation of quantum source *block* coding [10], [5], [1] for N signals requires $O(N^3)$ computational steps [2], a parallel implementation of quantum Huffman coding has only $O((\log N)^a)$ depth for some positive integer a, and a sequential implementation still uses just $O(N(\log N)^a)$ gates.

We will now describe our coding scheme for the storage of quantum signals. As before, we consider a quantum source emitting a sequence of independent and identically distributed quantum signals with a density matrix for each signal shown in (1) where p_i's are the eigenvalues. Considering Huffman coding for a classical source with probabilities p_i's allows one to construct a one-to-one correspondence between Huffman codewords h_i and the eigenstates $|\phi_i\rangle$. For parallel implementation, we find it useful to represent $|\phi_i\rangle$ by two pieces,[1] the first being the Huffman codeword, padded by the appropriate number of zeros to make it into constant length,[2] $|0\cdots 0h_i\rangle$, the second being the length of the Huffman codeword, $|l_i\rangle$, where $l_i = \text{length}\,(h_i)$. We also pad zeros to the second piece so that it becomes of fixed length $\lceil \log\, l_{\max}\rceil$ where $l_{\max}$ is the length of the longest Huffman codeword. Therefore, $|\phi_i\rangle$ is mapped into $|0\cdots 0h_i\rangle|l_i\rangle$. Notice that the length of the second tape is $\lceil \log\, l_{\max}\rceil$ which is generally small compared to n. The usage of the second tape is a small price to pay for efficient parallel implementation.

In this section, we use the model of a quantum gate array for quantum computation. The complexity class **QNC** is the class of quantum computations that can be performed in polylogarithmic parallel depth [7]. We prove the following theorem.

Theorem 1: Encoding or decoding of a quantum Huffman code for storage is in the complexity class **QNC.** Solving the classical Huffman coding problem for the eigenvalues of the density matrix gives a coding scheme with average codeword length $\overline{L}$ and maximum codeword length $l_{\max}$. For any $\delta > 0$, for large enough N, the quantum Huffman code stores data using less than $N(\overline{L}+\delta+\lceil \log\, l_{\max}\rceil)$ qubits. The encoding network has depth $O((\log N)^2)$.

The proof follows in the next two subsections.

A. Encoding

Without much loss of generality, we suppose that the total number of messages is $N = 2^r$ for some positive integer r. We propose to encode by divide and conquer. First, we divide the messages into pairs and apply a merging procedure to be discussed in (2) to each pair. The merging effectively reduces the total number of messages to 2^{r-1}. We can repeat this process. Therefore, after r applications of the merging procedure below, we obtain a single tape containing all the messages (in addition to the various length tapes containing the length information).

[1]The second piece contains no new information. However, it is useful for a massively parallel implementation of the shifting operations, which is an important component in our construction.

[2]The encoding process to be discussed below will allow us to reduce the total length needed for N signals.

The first step is the merging of two signals into a single message. Let us introduce a message tape. For simplicity, we simply denote $|0\cdots 0h_{i_1}\rangle$ by $|h_1\rangle$, etc.,

$$\begin{array}{ll} & |h_1\rangle|l_1\rangle|h_2\rangle|l_2\rangle \quad |\mathbf{0}\rangle_{\text{tape}} \\ \xrightarrow{\text{swap}} & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle \quad |0\cdots 0h_1\rangle_{\text{tape}} \\ \xrightarrow{\text{shift}} & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle \quad |h_1 0\cdots 0\rangle_{\text{tape}} \\ \xrightarrow{\text{swap}} & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle \quad |h_1 0\cdots 0h_2\rangle_{\text{tape}} \\ \xrightarrow{\text{shift}} & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle \quad |h_1 h_2 0\cdots 0\rangle_{\text{tape}}. \end{array} \tag{2}$$

We remark that the swap operation between any two qubits can be done efficiently by using an array of three XOR's with the two qubits alternately used as the control and the target.[3] The shift operation is just a permutation and therefore can be done in constant depth [7]. However, we actually need something slightly stronger: a controlled shift, controlled by functions of the lengths $|l_1\rangle$ and $|l_2\rangle$, which are quantum variables. To do a shift controlled by the register $|s\rangle$, we expand s in binary, and perform a shift by 2^i positions conditioned on the appropriate bit of s. When $|s\rangle$ is a quantum register in a superposition, this operation performed coherently will entangle the register with the tape, just as in the third difficulty described above. It is no longer a problem here, since we will disentangle the register and the tape during decoding.

Now the encoder keeps the original length tape for *each* signal as well as the message tape for two messages, i.e.,

$$|l_1\rangle|l_2\rangle|h_1 h_2 0\cdots 0\rangle_{\text{tape}}.$$

Notice that it is relatively fast to compute the length l_1+l_2 of the two messages from l_1 and l_2—$O(\log\, l)$ steps for the obvious sequential method (where l is the larger of l_1 and l_2), and $O(\log\log\, l)$ depth with a good parallel algorithm. Therefore, the merging procedure can be performed in polylogarithmic depth.

More concretely, at the end the encoder obtains

$$|l_1\rangle|l_2\rangle\cdots|l_N\rangle|h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}. \tag{3}$$

He has performed $\lceil \log\, N\rceil$ merges. Merging two messages of maximum length l requires $\lceil \log\, l\rceil$ shifts (each of constant depth) plus swaps (of constant depth) and one addition (of depth $O(\log\log\, l)$). The maximum length $l = Nl_{\max}$, so the full merging procedure requires depth $O((\log N)^2 + \log N \log l_{\max}$. In addition, there is a constant depth cost for performing the initial encoding, which we neglect in the large-N limit. We will also neglect the $\log N \log l_{\max}$ term.

Finally, the encoder truncates the message tape: He keeps only say the first $N(\overline{L}+\delta)$ qubits in the message tape $|h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}$ for some $\delta > 0$ and throws away the other qubits. This truncation minimizes the number of qubits needed. The only overhead cost compared to the classical case is the storage of the length tapes of the individual signals. This takes only $N\lceil \log\, l_{\max}\rceil$ qubits.[4]

B. Decoding

Decoding can be done by adding an appropriate number of qubits in the zero state $|0\rangle$ behind the truncated message tape and simply running the encoding process backward (again with only depth $O((\log N)^a)$).

What about fidelity? The key observation is the following:

Definition 2: The typical subspace S_δ is the subspace where the first $N(\overline{L}+\delta)$ qubits are arbitrary, and any qubits beyond that are in the *fixed* state $|0\cdots 0\rangle$.

[3]In (2), we do not include the position of the head, since it is simply dependent on the sum of the message lengths and can be reset to 0 after the process is completed.

[4]Further optimization may be possible. For instance, if $\log\, l_{\max}$ is large, one can save storage space by repeating the procedure, i.e., one can now use quantum Huffman coding for the problem of storing the quantum signals $|l_i\rangle$'s.

Proposition 3: $\forall \epsilon,\ \delta > 0, \exists N_0 > 0$ such that $\forall N > N_0$, $F \geq 1-\epsilon$ where F is the fidelity between the true state ρ of the N quantum signals and the projection of ρ on the typical subspace S_δ in our quantum Huffman coding scheme.

Proof: The proof is identical to the case of Schumacher's noiseless quantum coding theorem [10], [5], [1].

Therefore, the truncation and subsequent replacement of the discarded portion by $|0\cdots 0\rangle$ still lead to a high fidelity in the decoding.

In conclusion, we have constructed an explicit parallel encoding and decoding scheme for the storage of N independent and identically distributed quantum signals that asymptotically has only $O((\log\ N)^a)$ depth and uses $N(\overline{L}+\delta+\lceil\log\ l_{\max}\rceil)$ qubits for storage where $\overline{L}$ is the average length of the Huffman coding for the classical coding problem for the set of probabilities given by the eigenvalues of the density matrix of each signal. Here δ can be any positive number and $l_{\max}$ is the length of the longest Huffman codeword.

Corollary 4: A sequential implementation of the encoding algorithm requires only $O(N(\log\ N)^a)$ gates.

Proof: This follows immediately from the fact that the encoding is in **QNC** and uses $O(N)$ qubits: At each time step of a parallel implementation, only $O(N)$ steps are implemented. Since the network has depth $O((\log\ N)^a)$, there can be at most $O(N(\log\ N)^a)$ gates in the network.

IV. Communication

We now attempt to use the quantum Huffman coding for communication rather than for the storage of quantum signals. By communication, we assume that Alice receives the signals *one by one* from a source and is compelled to encode them one by one. As we will show below, the number of qubits required is slightly more, namely, $N(\overline{L}+\delta+\lceil\log\ l_{\max}\rceil)+\lceil\log\,(Nl_{\max})\rceil$. The code that we will construct is not instantaneous, but Alice and Bob can pay a small penalty in stopping the transmission any time. In fact, we have the following theorem.

Theorem 5: Sequential encoding and decoding of a quantum Huffman code for communication requires $N(\overline{L}+\delta+\lceil\log\ l_{\max}\rceil)+\lceil\log\,(Nl_{\max})\rceil$ qubits and only $O(N^2(\log\ N)^a)$ computational gates.

The proof follows in the next three subsections.

A. Encoding

The encoding algorithm is similar to that of Section III except that the signals are encoded one by one. More concretely, it is done through alternating applications of the swap-and-shift operations.

$$
\begin{aligned}
& |h_1\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|\mathbf{0}\rangle_{\text{tape}} \\
& \otimes|0\rangle_{\text{total length}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|0\cdots 0h_1\rangle_{\text{tape}} \\
& \otimes|0\rangle_{\text{total length}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|h_1 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|0\rangle_{\text{total length}} \\
\xrightarrow{\text{add}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|h_1 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|l_1\rangle_{\text{total length}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|h_1 0\cdots 0h_2\rangle_{\text{tape}} \\
& \otimes|l_1\rangle_{\text{total length}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|h_1h_2 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|l_1\rangle_{\text{total length}} \\
\xrightarrow{\text{add}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|h_1h_2 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|l_1+l_2\rangle_{\text{total length}} \\
& \cdots \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|\mathbf{0}\rangle|l_N\rangle|h_1h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|l_1+\cdots+l_{N-1}\rangle_{\text{total length}} \\
\xrightarrow{\text{add}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|\mathbf{0}\rangle|l_N\rangle|h_1h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}} \\
& \otimes|l_1+\cdots+l_N\rangle_{\text{total length}}.
\end{aligned}
\tag{4}
$$

We have included an ancillary space storing the total length of the codewords generated so far.[5] This space requires $\log(Nl_{\max})$ qubits.

Even though the encoding of signals themselves are done one by one, the shifting operation can be sped up by parallel computation. Indeed, as before, the required controlled-shifting operation can be performed in $O(\log\ N+\log\ l_{\max})$ depth. As before, if a sequential implementation is used instead, the complete encoding of one signal still requires only $O(N(\log\ N)^a)$ gates.

Now the encoding of the N signals in quantum communication is done sequentially, implying $O(N)$ applications of the shifting operation. Therefore, with a parallel implementation of the shifting operation, the whole process has depth $O(N(\log\ N)^a)$. With a sequential implementation, it takes $O(N^2(\log\ N)^a)$ steps.

B. Transmission

Notice that the message is written on the message tape from left to right. Moreover, starting from left to right, the state of each qubit once written remains unchanged throughout the encoding process. This decoupling effect suggests that rather than waiting for the completion of the whole encoding process, the sender, Alice, can start the transmission immediately after the encoding. For instance, after encoding the first r signals, Alice is absolutely sure that at least the first $rl_{\min}$ (where $l_{\min}$ is the minimal length of each codeword) qubits on the tape have already been written. She is free to send those qubits to Bob immediately. There is no penalty for such a transmission because it is easy to see that the remaining encoding process requires no help from Bob at all. (Note that in the asymptotic limit of large r, after encoding r signals, Alice can even send $r(\overline{L}-\epsilon)$ qubits for any $\epsilon > 0$ to Bob without worrying about fidelity.)

In addition, Alice can send the first r length variables $l_1, \cdots, l_r$, but she must retain the total-length variable for continued encoding. Since the total-length variable is entangled with each branch of the encoded state, decoding cannot be completed by Bob without use of this information. In other words, Alice must disentangle her system from the encoded message before decoding may be completed.

C. Decoding

With the length information of each signal and the received qubits, Bob can *start* the decoding process before the whole transmission is complete *provided that* he does not perform any measurement at this moment. For instance, having received $rl_{\min}$ qubits in the message tape from Alice, Bob is sure that at least $s = \lfloor rl_{\min}/l_{\max}\rfloor$ signals have already arrived. He can separate those s signals immediately using the length information of each signal. This part of the decoding process is rather straightforward and we will skip its description here.

The important observation is, however, the following: If Bob were to perform a measurement on his signals now, he would find that they are

[5] As in (2), we do not include the position of the head.

of poor fidelity. The reason behind this has already been noted in Section II. Even though the subsequent encoding process does not involve Bob's system, there is still entanglement between Alice and Bob's systems. More specifically, the shifting operations in the remaining encoding process by Alice require explicitly the information on the total length of decoded signals. Before Bob performs any measurement on his signals, it is, therefore, crucial for Alice to disentangle her system first, as mentioned above.

Suppose in the middle of their communication in which Bob has already received $K\overline{L}$ qubits from Alice, Bob suddenly would like to perform a measurement on his signals. He shall first inform Alice of his intention. Afterwards, one way to proceed is the following: They choose some convenient point, say the mth signal, to stop and consider quantum Huffman coding for only the first m signals and complete the encoding and decoding processes.

We shall consider two subcases. In the first subcase, the number m is chosen such that the mth signal is most likely still in the sender (Alice)'s hands (e.g., $m > K + O(\sqrt{K})$ in the asymptotic limit). The sender Alice now disentangles the remaining signal from the first m quantum signals by applying a quantum shifting operation. She can now complete the encoding process for quantum Huffman coding of the m signals and send Bob any untransmitted qubits on the tape. In the asymptotic limit of large K, $O(\sqrt{m})$ qubits of forward transmission (from Alice to Bob) are needed. (The required depth of the network is polynomial in $\log\ m$ if a parallel implementation of a quantum gate array is used.) In addition, Alice must send her record of the total length of the signals. However, this requires only an additional $\lceil \log\,(ml_{\max})\rceil$ qubits, so the total number which must be transmitted for disentanglement is still $O(\sqrt{m})$.

In the second subcase, the number m is chosen such that the mth signal is most likely already in the receiver (Bob)'s hands (e.g., $m < K - O(\sqrt{K})$ in the asymptotic limit). The receiver Bob now attempts to disentangle the remaining signals from the first m quantum signals by applying a quantum shifting operation. Of course, he needs to shift some of his qubits back to Alice. This asymptotically amounts to $O(\sqrt{m})$ qubits of *backward* communication. This is a penalty that one must pay for this method. After this is done, Alice must again send her length register to Bob (after subtracting the lengths of the signals returned to her). This requires an additional $O(\log\ m)$ qubits.

If m is chosen between $K - O(\sqrt{K})$ and $K + O(\sqrt{K})$, neither sending signals forward or backward will suffice to properly disentangle the varying lengths of the signals. One possible solution is to choose $m' > K + O(\sqrt{K})$ and perform the above procedure, sending m' total signals to Bob. Then Bob decodes and returns the $m' - m$ extra signals to Alice. This method requires $O(\sqrt{K})$ qubits transmitted forward and $O(\sqrt{K})$ qubits transmitted backward to disentangle.

We remark that the shifting operation can be done rather easily in distributed quantum computation between Alice and Bob. This is a nontrivial observation because the number of qubits to be shifted from Alice to Bob is itself a quantum-mechanical variable. This, however, does not create much problem. Bob can always communicate with Alice using a bus of fixed length. For example, he applies local operations to swap the desired quantum superposition of various numbers of qubits from his tape to the bus, sends such a bus to Alice, etc.

The result is the following theorem.

Theorem 6: Alice and Bob may truncate a communication session after the transmission of m encoded signals, retaining high fidelity with the cost of $O(\sqrt{m})$ additional qubits transmitted.

In the above discussion, we have focused on the simple case when Bob would like to perform a measurement on the whole set of the first m signals. Suppose Bob is interested only in a particular signal, say the mth one, but not the others. There exists a more efficient scheme for doing it. We shall skip the discussion here.

V. Concluding Remarks

We have successfully constructed a Huffman-coding-inspired scheme for the storage of quantum information. Our scheme is highly efficient. The encoding and decoding processes of N quantum signals can be done *in parallel* with depth polynomial in $\log\ N$. (If parallel machines are unavailable, as shown in Section IV-A our encoding scheme will still take only $O(N(\log\ N)^a)$ computational steps for a sequential implementation. In contrast, a naive implementation of Schumacher's scheme will require $O(N^3)$ computational steps.) This massive parallelism is possible because we explicitly use another tape to store the length information of the individual signals. The storage space needed is asymptotically $N(\overline{L} + \delta + \lceil \log\ l_{\max}\rceil)$ where $\overline{L}$ is the average length of the corresponding classical Huffman coding problem for the density matrix in the diagonal form, δ is an arbitrary small positive number, and $l_{\max}$ is the length of the longest Huffman codeword.

We also considered the problem of using quantum Huffman coding for communication in which case Alice encodes the signals one by one. $N(\overline{L} + \delta + \lceil \log\ l_{\max}\rceil) + O(\log\ N)$ qubits are needed. With a parallel implementation of the shifting operation, depth of $O(N(\log\ N)^a)$ is needed. On the other hand, with a sequential implementation, $O(N^2(\log\ N)^a)$ computational steps are needed. In either case, the code is not instantaneous, but, by paying a small penalty in terms of communication and computational costs, Alice and Bob have the option of stopping the transmission and Bob may then start measuring his signals.

More specifically, while the receiver Bob is free to separate the signals from one another, he is not allowed to measure them until the sender Alice has completed the encoding process. This is because Alice's encoder head generally contains the information of the total length of the signals. In other words, its state is entangled with Bob's signals. Therefore, whenever Bob would like to perform a measurement, he should first inform Alice and the two should proceed with disentanglement. We present two alternative methods of achieving such disentanglement one of which involves forward communication and the other of which involves both forward and backward.

Since real communication channels are always noisy, in actual implementation source coding is always followed by encoding into an error-correcting code. Following the pioneering work by Shor [11] and independently by Steane [12], various quantum error-correcting codes have been constructed. We remark that quantum Huffman coding algorithm (even the version for communication) can be immediately combined with the encoding process of a quantum error-correcting code for efficient communication through a noisy channel.

As quantum information is fragile against noises in the environment, it may be useful to work out a fault-tolerant procedure for quantum source coding. The generalizations of other classical coding schemes to the quantum case are also interesting [6]. Moreover, there exist universal quantum data compression schemes motivated by the Lempel–Ziv compression algorithm for classical information [8].

Acknowledgment

H.-K. Lo would like to thank D. P. DiVincenzo, J. Preskill, and T. Spiller for helpful discussions.

References

[1] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, "General fidelity limit for quantum channels," *Phys. Rev.*, vol. A54, p. 4707, 1996.

[2] R. Cleve and D. P. DiVincenzo, "Schumacher's quantum data compression as a quantum computation," *Phys. Rev.*, vol. A54, p. 2636, 1996.
[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[4] R. Jozsa, "Fidelity for mixed quantum states," *J. Mod. Opt.*, vol. 41, p. 2315, 1994.
[5] R. Jozsa and B. Schumacher, "A new proof of the quantum noisless coding theorem," *J. Mod. Opt.*, vol. 41, p. 2343, 1994.
[6] R. Jozsa, M. Horodecki, P. Horodecki, and R. Horodecki, "Universal quantum information compression," *Phys. Rev. Lett.*, vol. 81, p. 1714, 1998.
[7] C. Moore and M. Nilsson. Parallel Quantum Computation and Quantum Codes (Los Alamos e-print archive). [Online] Available: http://xxx.lanl.gov/abs/quant-ph/9808027
[8] M. A. Nielsen, "Quantum information theory," Ph.D. dissertation, Univ. New Mexico, Albuquerque, 1998.
[9] B. Schumacher, "Quantum Kraft Inequality," presented at the Santa Fe Institute, 1994.
[10] ——, "Quantom coding," *Phys. Rev.*, vol. A51, p. 2738, 1995.
[11] P. W. Shor, "Scheme for reducing decoherence in quantum computer memory," *Phys. Rev.*, vol. A52, p. R2493, 1995.
[12] A. M. Steane, "Error correcting codes in quantum theory," *Phys. Rev. Lett.*, vol. 77, p. 793, 1996.

Optimization of Distributed Detection Systems Under the Minimum Average Misclassification Risk Criterion

Maurizio Magarini and Arnaldo Spalvieri

***Abstract*—A common model for distributed detection systems is that of several separated sensors each of which measures some observable, quantizes it, and communicates to a fusion center the quantized observation. The fusion center collects the quantized observations and takes the decision. The present correspondence deals with the design of the quantizers and of the fusion center under a rate constraint. The system of interest allows soft nonbreakpoint quantizers and nonindependent observations. Our finding is that locally optimal design of the distributed detection system is feasible via alternate minimization of the average misclassification risk.**

***Index Terms*—Alternate optimization, average misclassification risk, distributed detection.**

I. Introduction

Distributed detection systems have received a lot of attention in the past two decades, as documented in the special issue of the PROCEEDINGS OF THE IEEE [1]. A common model for these systems involves several separated sensors, each of which measures some observable, quantizes it, and communicates to a fusion center the quantized observation. The fusion center collects the quantized observations and takes the decision. Since the rate of transmission between the sensors and the fusion center is a cost, fine quantization of data may be not allowed. A crucial problem is therefore the design of coarse quantizers that satisfy a rate constraint and that introduce low degradation in the detection capability of the system. Tsitsiklis and Athans have shown in [2] that, when conditional independence of the observation given the hypothesis cannot be assumed, the design problem is NP complete. Hence one is lead to renounce to global optimality and to study suboptimal strategies. Several design strategies have been studied in the past, most of which were tailored to hard (one-bit) quantizers. Tenney and Sandell optimized the decentralized quantizers with a fixed fusion rule [3], while Chair and Varshney considered the design of the fusion rule for fixed quantizers [4]. Joint design of soft (multibit) quantizers has been studied by Longo *et al.*in [5], where an alternate optimization technique is proposed. Specifically, the approach in [5] is to maximize the Bhattacharyya distance between the multivariate conditional probabilities of quantized data given the hypotheses. The potential weakness of this approach is that the Bhattacharyya distance is not the natural measure of performance of detection systems. Therefore, one wonders whether joint design of quantizers and the fusion rule under the natural criterion of performance is feasible. Our answer is that locally optimal design, that is, minimization of the average misclassification risk, is feasible by alternate optimization. A similar method was adopted in [6] in the framework of decentralized parameter estimation. Also, in [7] the alternate optimization technique is considered as a method to minimize a general distortion measure. Like [5], [7], our method applies to nonindependent observations and to soft (multibit) nonbreakpoint quantizers.

Manuscript received November 19, 1998; revised December 22, 1999.

The authors are with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133 Milano, Italy (e-mail: magarini@elet.polimi.it; spalvier@elet.polimi.it).

Communicated by P. A. Chou, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(00)05016-1.

II. System Model and Problem Statement

For the sake of simplicity, consider two scalar observations and binary detection. Extensions are straightforward. Let x_1, x_2 denote the observations, and assume that they are drawn from the continuous spaces $\mathcal{X}_1$, $\mathcal{X}_2$. In the classical formulation of the detection problem, a hidden discrete random variable (the *class*, or the *hypothesis*) is drawn together with the observation vector according to some known joint probability distribution. We call such a discrete random variable $c \in \mathcal{C} = \{c_1, c_2\}$. The goal of the detection system is to guess the hidden class given the observation vector.

A. System Description

The decentralized detection system we are concerned with is modeled as a decision rule made by two scalar quantizers and a fusion center. Each scalar quantizer is allowed here to be a nonbreakpoint one. Quantizer $Q_n(x_n)$, $n = 1, 2$, is modeled as a mapping from $\mathcal{X}_n$ to $\mathcal{I}_n$, where $\mathcal{I}_n = \{0, 1, \cdots, I_n - 1\}$. Of course, the rate R_n of the nth quantizer is $R_n = \log_2 I_n$. Inversion of $Q_n(x)$ is hereafter intended as

$$Q_n^{-1}(i) = \{x_n \in \mathcal{X}_n : Q_n(x_n) = i\}.$$

The decision function performed by the fusion center, denoted $\Phi(i_1, i_2)$, is a mapping from $\mathcal{I}_1 \times \mathcal{I}_2$ to $\mathcal{C}$. The decision rule of the decentralized detection system, denoted $\Phi(Q_1(x_1), Q_2(x_2))$, is a mapping from $\mathcal{X}_1 \times \mathcal{X}_2$ to $\mathcal{C}$. As in [5], we assume that the processing to be performed at the fusion center is unlimited in complexity. In practice, this means that the fusion center is a lookup table with $2^{R_1+R_2}$ entries. A pictorial example of the decision rule for a specific two-dimensional decentralized detection system is later illustrated in Fig. 6.

B. Statement of the Problem

The Bayesian risk (or cost) in deciding in favor of class $\hat{c} \in \mathcal{C}$ when x_1, x_2 is observed is

$$R(\hat{c}|x_1, x_2) = \sum_{i=1}^{2} b(c_i \mapsto \hat{c}) P(c_i|x_1, x_2) \tag{1}$$