Optimal and efficient decoding of concatenated quantum block codes

David Poulin
Center for the Physics of Information, California Institute of Technology, Pasadena, California 91125, USA
(Received 23 June 2006; published 22 November 2006)

We consider the problem of optimally decoding a quantum error correction code—that is, to find the optimal recovery procedure given the outcomes of partial "check" measurements on the system. In general, this problem is NP hard. However, we demonstrate that for concatenated block codes, the optimal decoding can be efficiently computed using a message-passing algorithm. We compare the performance of the message-passing algorithm to that of the widespread blockwise hard decoding technique. Our Monte Carlo results using the five-qubit and Steane's code on a depolarizing channel demonstrate significant advantages of the message-passing algorithms in two respects: (i) Optimal decoding increases by as much as 94% the error threshold below which the error correction procedure can be used to reliably send information over a noisy channel; and (ii) for noise levels below these thresholds, the probability of error after optimal decoding is suppressed at a significantly higher rate, leading to a substantial reduction of the error correction overhead.

DOI: 10.1103/PhysRevA.74.052333 PACS number(s): 03.67.Pp, 03.67.Hk, 03.67.Lx

I. INTRODUCTION

Quantum error correction (QEC) [1] and fault-tolerant quantum computation [2] demonstrate that quantum information can, in principle, be stored and manipulated coherently for arbitrarily long times despite the presence of noise. The general framework of QEC is the following. Redundancy is introduced by encoding the information of system S into a larger system S'. The image of S in S' characterizes a code, while a particular embedding of S into S' is called an encoding. The system S' is subjected to some noise. Partial measurements whose outcomes are known as the "error syndrome" are performed on S'. Conditioned on this error syndrome, a recovery operation is applied to S' in order to restore its original information. This last step, called "decoding," is the subject of the present study.

In the absence of structure in the code, we know from a classical result [3] that finding the optimal recovery is NP hard. For practical purposes, one must either use codes with lots of structure, which typically offer poorer performances, or settle for suboptimal recovery. Residual errors after decoding are therefore of two varieties: those due to the information-theoretic limitations of the code and those arising from suboptimal decoding procedures. In the past decades, considerable progress has been made towards understanding this trade-off in the classical setting (see, e.g., [4,5], and references therein). Central to these advancements is the use of the message-passing decoding algorithm pioneered by Gallager [6], which often leads to near-optimal decoding. This technique was recently introduced in the quantum realm by Ollivier and Tillich [7] and Camara *et al.* [8] for the decoding of low-density parity check (LDPC) codes (see also [9,10], for related work).

Concatenation of block codes is widely used in quantum information science and a key component of almost all fault tolerant schemes (a noticeable exception is topological quantum computing [11]). As the name suggests, the system S' that redundantly encodes the information of system S can itself be encoded in a yet larger system S'', adding an extra layer of redundancy. Provided the initial error rate is below a threshold value, every extra level of concatenation should reduce the probability of error after decoding, so concatenation can, in principle, be repeated until the error is below any desired value.

In this paper, we demonstrate an efficient [25] message-passing algorithm that achieves optimal (maximum likelihood) decoding for *concatenated block codes* with uncorrelated noise. We numerically investigate the message-passing algorithm using the five-qubit code [12] and Steane's seven-qubit code [13] and compare their performances to the commonly used blockwise minimal-distance decoder (based on a local rather than global optimization). The advantages of the message-passing algorithm are substantial. On the one hand, for the five-qubit code used on a depolarizing channel, the message-passing algorithm can correctly decode the information for a noise level up to at least 0.1885 (the exact threshold is probably the hashing bound ≈ 0.189) compared to the values 0.1376 previously established using blockwise decoding [14]. For Steane's code, this enhancement is even greater going from 0.0969 [14] to at least 0.188. On the other hand, away from these noise thresholds, the probability of error decreases at a significantly higher rate using optimal decoding. For instance, for a 0.1 depolarizing channel and using four levels of concatenation of the five-qubit code, the probabilities that the blockwise decoding and the optimal decoding fail to correctly identify the error differ by more that three orders of magnitude. As a consequence, a decoding error probability $p_e \leq \delta$ for any $\delta > 0$ can be achieved with a substantially reduced error correction overhead.

II. STABILIZER FORMALISM

Our presentation of the stabilizer formalism follows [15], see [16] for the general theory. Denote by X, Y, and Z, the three Pauli matrices and by $\mathbb{1}$ the 2×2 identity matrix. The group $\mathcal{P}_1$ is the multiplicative group generated by the Pauli matrices and the imaginary unit i. The n-qubit Pauli group $\mathcal{P}_n$ is the n-fold tensor product of $\mathcal{P}_1$. We denote X_j the Pauli

matrix X acting on the jth qubit for $j=1,\ldots,n$ and similarly for Y and Z. Note that the X_j's and the Z_j's are a generating set of $\mathcal{P}_n$, i.e., $\mathcal{P}_n=\langle i,X_j,Z_j\rangle$. The Clifford group on n qubits $\mathcal{C}_n$ is the largest subgroup of the unitary group $U(2^n)$ that maps $\mathcal{P}_n$ to itself under the adjoint action.

The encoding of k qubits into n qubits can be specified by a matrix $C\in\mathcal{C}_n$. C is a unitary matrix acting on n qubits that are distributed in three different sets. The first k "logical" qubits contain the information to be encoded in the n qubits; the next u "stabilizer" qubits are set to the state $|0\rangle^{\otimes u}$; and finally the remaining $r=n-k-u$ "gauge" qubits are in arbitrary states. The image of the Pauli operators acting on the first k qubits are known as logical Pauli operators $\bar{X}_j=CX_jC^\dagger$ and $\bar{Z}_j=CZ_jC^\dagger$. The image of the Z Pauli operators acting on qubits $j=k+1,\ldots,k+u$ are called stabilizer generators $S_j=CZ_{k+j}C^\dagger$, whereas the image of the X operators acting on those qubits are called pure errors $T_j=CX_{k+j}C^\dagger$. Finally, the image of the Pauli operators acting on the remaining r qubits are called gauge operators $g_j^x=CX_{k+u+j}C^\dagger$ and $g_j^z=CZ_{k+u+j}C^\dagger$.

The stabilizer generators S_j mutually commute, so can be simultaneously measured. The outcome of that measurement is called the error syndrome $s\in\{-1,1\}^u$. Since the u stabilizer qubits are all in state $|0\rangle$ prior to encoding, we conclude that in the absence of noise the encoded state should be a $+1$ eigenstate of all stabilizer generators; thus, the error syndrome should be all ones. A nontrivial syndrome therefore indicates that an error has corrupted the register, and the task of decoding consists in finding the optimal recovery procedure given an error syndrome.

III. DECODING

To address the decoding problem, note that $\mathcal{P}_n=\langle i,\bar{X}_j,\bar{Z}_j,S_j,T_j,g_j^x,g_j^z\rangle$. In other words, any element $E\in\mathcal{P}_n$ can be written, up to an irrelevant phase, as

$$E=\mathcal{L}(E)\mathcal{T}(E)\mathcal{G}(E), \tag{1}$$

where $\mathcal{L}(E)$ is a product of logical Pauli operators, $\mathcal{T}(E)$ is a product of pure errors, and $\mathcal{G}(E)$ is a product of gauge operators and stabilizer elements. Moreover, this decomposition can be found by running the circuit C backward, which is efficient since $C\in\mathcal{C}_n$ [16]. $\mathcal{T}(E)$ is completely determined by the syndrome: T_j appears in $\mathcal{T}(E)$ if and only if the jth syndrome bit is -1. The value of $\mathcal{G}(E)$ is irrelevant because the information encoded in the n qubits is invariant under the action of any $\mathcal{G}(E)$. This reflects the fact that the stabilizer qubits are initially set to $|0\rangle$ and that the gauge qubits are in random states. Thus, to undo the effect of an error E, one needs to identify the most likely value of $L=\mathcal{L}(E)$ given s, or equivalently given $T=\mathcal{T}(E)$.

For simplicity, we will focus on Pauli channels, where errors E are elements of $\mathcal{P}_n$ distributed according to $P(E)$. Given this probability $P(E)$ over $\mathcal{P}_n$, one can compute the conditional probability $P(L|T)=P(L,T)/P(T)$ using

$$P(L,T)=\sum_E \delta[\mathcal{T}(E)=T]\delta[\mathcal{L}(E)=L]P(E) \tag{2}$$

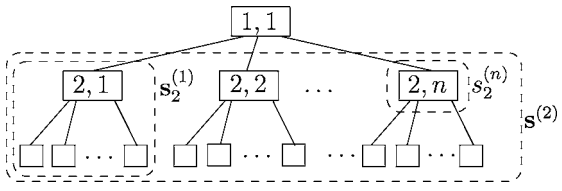
$$=\sum_G P(E=LTG), \tag{3}$$

where δ denotes the indicator function and G takes all possible combinations of stabilizer generators and gauge operators. Given a finite block size n, these probabilities can be computed and the optimal decoding $\hat{L}(T)=\mathrm{argmax}_L\{P(L|T)\}$ can be evaluated. Decoding a block code thus consists of looking in a table containing the values of $\hat{L}(T)$ for each T. Typically—and, in particular, for a nondegenerate code over the depolarization channel—$\hat{L}(T)$ corresponds to the minimal distance decoder $\mathcal{L}(\hat{E}(T))$, where $\hat{E}(T)$ is the error acting on the fewest number of qubits and that is compatible with the observed syndrome.

Concatenation is realized by encoding the n qubits of the code in another code. There is no need for this other code to be identical to the original one. However to simplify the presentation, we will assume that the same code is used at every concatenation layer and that it encodes a single qubit in n qubits; generalizations are straightforward. This procedure can be repeated ℓ times at the expense of an exponentially growing number of physical qubits n^ℓ. The number of stabilizer generators grows roughly as $un^{\ell-1}$ (it is a geometric sum); thus, the syndrome takes $2^{un^{\ell-1}}$ different values. Thus, even for moderate values of ℓ, it is not feasible to construct a lookup table giving the optimal decoding procedure for each syndrome value.

What is generally done to circumvent this double exponential blowup is to apply the optimal recovery *independently for each concatenation layer* (see, e.g., [16], Chap. 6 and references therein). One first measures the syndrome from each of the $n^{\ell-1}$ blocks of n qubits of the last layer of concatenation, and optimally decodes them using the lookup table. One then moves one layer up and applies the same procedure to the $n^{\ell-2}$ blocks of the second-to-last layer, etc. When the initial error rate is below a certain threshold value, the probability p_e that this procedure fails to correctly identify $\mathcal{L}(E)$ decreases doubly exponentially with ℓ. Hence, this decoding scheme based on hard decisions for each concatenation layer is efficient and leads to a good error suppression, but is nonetheless suboptimal.

Optimal decoding

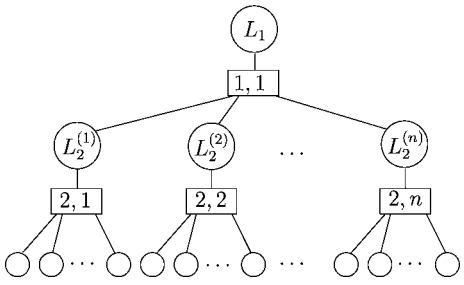


Let $s_m^{(j)}\in\{-1,1\}^u$ be the syndrome of the jth block of the mth concatenation layer. Denote $\mathbf{s}_m^{(j)}$ the collection of syndromes whose stabilizers act nontrivially on the physical qubits associated to the jth block of the mth concatenation

layer: these sets can be defined recursively by $\mathbf{s}_m^{(j)}=\{s_m^{(j)}\}\cup\{\cup_{i=jn-j+1}^{jn}\mathbf{s}_{m+1}^{(i)}\}$ with the initialization $\mathbf{s}_\ell^{(j)}=s_\ell^{(j)}$. Finally, denote $\mathbf{s}_m=\cup_{j=1}^{n^{m-1}}\mathbf{s}_m^{(j)}$ all the syndromes from the layers m to ℓ. (See the above diagram for a pictorial representation of $s_m^{(j)}$, $\mathbf{s}_m^{(j)}$, and $\mathbf{s}_m$.) Then, $\mathbf{s}_1$ is the set of all syndromes and maximum likelihood decoding consists in finding $\mathrm{argmax}_{L_1}P(L_1|\mathbf{s}_1)$. This probability can be factorized by conditioning on the logical errors of the second layer $\mathbf{L}_2=(L_2^{(1)},\ldots,L_2^{(n)})$,

$$\begin{aligned}P(L_1|\mathbf{s}_1)&=\sum_{\mathbf{L}_2}P(L_1|\mathbf{s}_1,\mathbf{L}_2)P(\mathbf{L}_2|\mathbf{s}_1)\\&=\sum_{\mathbf{L}_2}\delta[L_1=\mathcal{L}(\mathbf{L}_2)]\frac{P(\mathbf{L}_2,\mathbf{s}_1)}{P(\mathbf{s}_1)}\\&=\sum_{\mathbf{L}_2}\delta[L_1=\mathcal{L}(\mathbf{L}_2)]\frac{P(s_1|\mathbf{L}_2,\mathbf{s}_2)P(\mathbf{L}_2,\mathbf{s}_2)}{P(s_1,\mathbf{s}_2)}\\&=\sum_{\mathbf{L}_2}\delta[L_1=\mathcal{L}(\mathbf{L}_2)]\delta[s_1=\mathcal{S}(\mathbf{L}_2)]\frac{P(\mathbf{L}_2|\mathbf{s}_2)P(\mathbf{s}_2)}{P(s_1,\mathbf{s}_2)}\\&=\sum_{\mathbf{L}_2}\frac{\delta[L_1=\mathcal{L}(\mathbf{L}_2)]\delta[s_1=\mathcal{S}(\mathbf{L}_2)]}{P(s_1|\mathbf{s}_2)}\prod_{j=1}^{n}P(L_2^{(j)}|\mathbf{s}_2^{(j)}).\end{aligned}\tag{4}$$

Above, $\mathcal{S}(\mathbf{L})$ denotes the syndrome associated to the error pattern $\mathbf{L}\in\mathcal{P}_n$. This series of manipulations repeatedly uses Bayes' rule and the fact that the syndrome and logical error of level m are completely determined given the logical errors of layer $m+1$. The last step relies on the important assumption that the channel is memoryless, or more specifically, that the noise model does not correlate qubits across distinct blocks (errors on qubits in the same block could be correlated).



Equation (4) shows that by conditioning on the logical errors of each concatenation layer, the *factor graph* associated to the function $P(L_1|\mathbf{s}_1)$ is a tree, as depicted in the above graph. We have thus reduced optimal decoding to a SUM PRODUCT problem (known as tensor network contraction in quantum information science [17]) on a tree graph, which can be solved exactly and efficiently in the number of variables using a message passing algorithm (also known as belief propagation); see [4,5,18], and references therein. Let us describe this algorithm in a general setting.

The factor graph is a bipartite graph, and vertices from the two partitions are decorated with circles and boxes. Circle vertices are labeled $c=1,\ldots,N$ and each one carries a variable x_c with value in a discrete set. Box vertices are labeled $b=1,\ldots,M$, and each one contains a function f_b that depends on the variables x_c from the adjacent circles $c\in\mathcal{N}(b)$, collectively denoted $X_b=\{x_c:c\in\mathcal{N}(b)\}$. The goal is to compute marginals

$$f(x_c)=\frac{1}{Z}\sum_{\{x_1,\ldots x_N\}\backslash x_c}\prod_{b=1}^{M}f_b(X_b),\tag{5}$$

where $\backslash x_c$ indicates that x_c is omitted from the set and Z is a normalization factor. To this end, messages $q_{c\to b}$ are passed from the circles to the boxes and messages $r_{b\to c}$ are passed from the boxes to the circles following the rules

$$q_{c\to b}(x_c)=\prod_{b'\in\mathcal{N}(c)\backslash b}r_{b'\to c}(x_c),\tag{6}$$

$$r_{b\to c}(x_c)=\sum_{X_b\backslash x_c}\left[f_b(X_b)\prod_{c'\in\mathcal{N}(b)\backslash c}q_{c'\to b}(x_{c'})\right],\tag{7}$$

where $\mathcal{N}(b)\backslash c$ means all neighbors of b excluding c, and similarly for $\mathcal{N}(c)\backslash b$. Note that these messages are functions of the discrete variables x_c (i.e., they are arrays). The $q_{c\to b}$ messages are initialized to the constant function 1. For a tree graph, the desired marginal is obtained from these messages after a number of steps equal to the depth of the variable x_c and is given by $f(x_c)=k\Pi_{b\in\mathcal{N}(c)}r_{b\to c}$, where k is a normalization factor.

In the case of interest, circles carry logical operators and a box labeled m,j carries the function $\delta[L_{m+1}^{(j)}=\mathcal{L}(L_m^{(nj-n+1)},\ldots,L_m^{(nj)})]\delta[s_{m+1}^{(j)}=\mathcal{S}(L_m^{(nj-n+1)},\ldots,L_m^{(nj)})]$, where the syndrome is fixed by the measurements. To complete the picture, extra box vertices carrying the function Eq. (3) need to be attached to the bottom leaves of the graph. The factor $p(s_1|\mathbf{s}_2)^{-1}$ can be evaluated by normalizing the obtained distribution. Thus, we can efficiently evaluate $P(L_1|\mathbf{s}_1)$ [26], and the optimal recovery is the L_1 maximizing this function.

The advantage of the message-passing algorithm over the minimal distance decoder comes from the fact that it does not throw away useful information [19]. Instead of computing the most likely recovery and passing it on to the next level of coding, the entire list of probability of possible recoveries, conditioned on the observed syndrome, is passed on. In other words, the original channel is composed with the syndrome measurement, and projected onto the logical algebra to yield a "conditionally renormalized" channel.

IV. NUMERICAL RESULTS

Following the tradition for benchmarking QEC techniques, we investigate the performance of the message-passing decoding algorithm using a depolarization channel, where each qubit is independently subjected to the channel

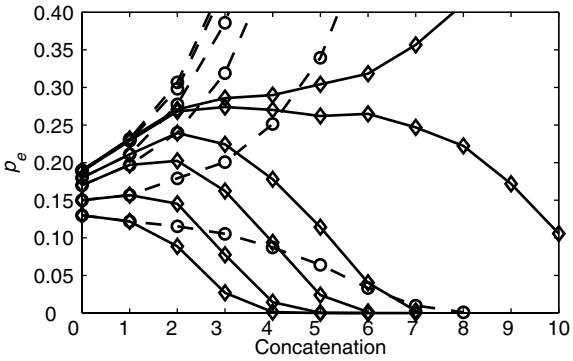


FIG. 1. Monte Carlo results for the five-qubit code showing the probability of erroneous decoding p_e as a function of the level of concatenation ℓ for different depolarization rate $p=0.13$, 0.15, 0.17, 0.18, 0.1885, and 0.19. The diamonds are from the message-passing algorithm, and the circles are from the blockwise decoding. All data are from samples of 2×10^4 encoded qubits.

$$\mathcal{E}_p(\rho) = (1-p)\rho + \frac{p}{3}(X\rho X + Y\rho Y + Z\rho Z). \tag{8}$$

We use the five-qubit code [12] concatenated with itself up to $\ell=10$ times, for an overhead of 9 765 625 physical qubits per logical qubit. Pauli errors $E\in\mathcal{P}_{n^\ell}$ are generated by picking each n single-qubit operator independently according to the probability $P(\mathbb{1})=1-p$, $P(X)=P(Y)=P(Z)=p/3$. The associated logical error $\mathcal{L}(E)$ and syndromes $\mathcal{S}(E)$ are computed exactly. These syndromes are used by a blockwise decoding routine yielding an estimate L_{BW} and by a message-passing routine yielding the optimal decoding $\hat{L}$. A decoding is declared incorrect when its estimate differs from $\mathcal{L}(E)$. This is repeated a large number of times (10^4-10^8) to evaluate the probability p_e that the decoding gives an incorrect estimate.

Figure 1 shows the probability of incorrect decoding p_e for both the blockwise and the optimal decoding as a function of the level of concatenation ℓ and for different channel parameters p ranging from 0.13 to 0.19. For the blockwise decoding, p_e ceases to decrease with ℓ for values of $p\geq0.15$. This reflects the fact that the threshold of this decoding technique for this particular code is about 0.1376 [14], so all curves except the 0.13 one are above the threshold. On the other hand, optimal decoding succeeds in decreasing the error probability for values of p up to at least 0.1885, but appears to fail at $p=0.19$. We conjecture that the exact value of this threshold is the hashing bound $\simeq0.189$, where the single-qubit coherent information vanishes and is the highest threshold any nondegenerate code can achieve [20]. Results obtained from Steane's code [13] show a quite similar behavior, with at least 94% increase of threshold going from 0.0969 [20] to at least 0.188 and appears to fail at 0.1885.

An interesting feature of the $p_e(\ell)$ curves obtained from optimal decoding is their nonmonotonicity. Blockwise decoding, on the other hand, always yields monotonic curves for this type of channel; thus, its global behavior under concatenation can be predicted from a single level of coding. This is because decoding is performed independently on each concatenation layer. With the optimal decoder, information about the syndromes is propagated from one layer of concatenation to the next through the conditionally renormalized channel that ceases to be depolarizing and varies from one qubit to the other. Thus, nonmonotonicity of the $p_e(\ell)$ curves is a signature of the global optimization performed by the message-passing algorithm.

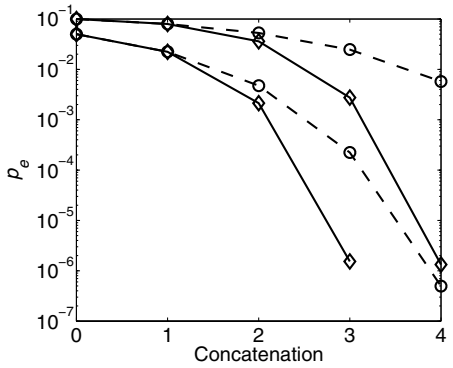


FIG. 2. As in Fig. 1 for $p=0.1$ and 0.05. Diamonds obtained from samples of 10^8 encoded qubits. Circles were produced using an exact numerical technique similar to that of Ref. [14].

Figure 2 shows the behavior of p_e as a function of ℓ away from the threshold values, i.e., in the natural operating regime of the code. Again, the advantages of the message-passing algorithm are considerable. After four rounds of concatenation for $p=0.1$, message passing fails with a probability of roughly 10^{-6}, whereas this probability is well above 10^{-3} for blockwise hard decoding. It takes six layers of concatenation for the blockwise decoding to reach comparable performances. Again, results obtained from Steane's code show an even larger gap.

Finally, we once again stress that the message passing outputs the probability of an error L rather than a particular value of L. A hard decision can then be made based on this probability. We observe that when decoding succeeds, $P(\hat{L})$ is typically very close to one (e.g., 0.999 for $\ell=3$), whereas when it fails it is relatively low (typically, 0.7); the algorithm knows that it is failing. This "flagging" of errors offers a great advantage when postselection is an option. The possibility of operating the algorithm with soft inputs, i.e., noisy syndrome measurements, is also of interest in several circumstances.

V. CONCLUSION

We have demonstrated an efficient message-passing algorithm for the optimal decoding of concatenated quantum block codes on a memoryless channels. Numerical results show substantial benefits of our approach over the widely used blockwise hard decoding, including an increase of error thresholds and a greater error suppression rate. Message-passing algorithms have been used on graphs with loops (describing, e.g., LDPC codes, turbo codes, or channels with memory) and often yield near-optimal decoding. The quantum generalization of these schemes, including quantum LDPC codes [8,9] and quantum turbo codes [22], are prom-

ising avenues for the realization of quantum information technologies. Techniques reminiscent of message passing have been used to beat the hashing bound but were not efficiently implementable [20,21]: efficient decoding may now be within reach using our techniques. A "hard" message-passing scheme was also used in [23] to obtain high fault-tolerant error thresholds: a full-fledge message-passing scheme, although not optimal for correlated errors that are typically present in fault-tolerant schemes, should further improve this threshold and may significantly reduce the resource overhead.

ACKNOWLEDGMENT

I thank Harold Ollivier for several useful conversations on message-passing algorithms and quantum error correction, and Graeme Smith and Jon Yard for comments. This work was supported, in part, by the Gordon and Betty Moore Foundation through Caltech's Center for the Physics of Information, by the National Science Foundation under Grant No. PHY-0456720, and by the National Science and Engineering Research Council of Canada.

[1] P. W. Shor, Phys. Rev. A **52**, R2493 (1995); A. M. Steane, Proc. R. Soc. London, Ser. A **452**, 2551 (1996); E. Knill and R. Laflamme, Phys. Rev. A **55**, 900 (1997).

[2] P. W. Shor, in *Proceedings of the 37th Symposium on the Foundations of Computer Science* (IEEE Press, Los Alamitos, CA, 1996), pp. 56–65; A. Y. Kitaev, in *Proceedings of the Third International Conference on Quantum Communication and Measurement*, edited by O. Hirota, A. S. Holevo, and C. M. Caves (Plenum, New York, 1997), pp. 181–188; E. Knill, R. Laflamme, and W. H. Zurek, Philos. Trans. R. Soc. London, Ser. A **454**, 365 (1998); D. Aharonov and M. Ben-Or, in *Proc. 29th Ann. ACM Symp. on Theory of Computing* (ACM, New York, 1997); J. Preskill, Proc. R. Soc. London, Ser. A **454**, 385 (1998).

[3] E. R. Berlekamp, R. J. McEliece, and H. van Tilborg, IEEE Trans. Inf. Theory **24**, 384 (1978).

[4] T. Richardson and R. Urbanke, *Modern Coding Theory* (to be published), available online at http://lthcwww.epfl.ch/mct/

[5] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).

[6] R. G. Gallager, *Low Density Parity Check Codes* (MIT Press, Cambridge, MA, 1963).

[7] H. Ollivier and J.-P. Tillich, in *Proceedings of 26th Symposium on Information Theory in the Benelux* (Brussels, 2005), p. 149.

[8] T. Camara, H. Ollivier, and J.-P. Tillich, e-print quant-ph/0502086.

[9] D. J. C. MacKay, G. Mitchison, and P. L. McFadden, IEEE Trans. Inf. Theory **50**, 2315 (2004).

[10] H. Ollivier and J.-P. Tillich, Phys. Rev. A **74**, 032304 (2006).

[11] A. Y. Kitaev, Ann. Phys. (N.Y.) **303**, 2 (2003).

[12] R. Laflamme, C. Miquel, J.-P. Paz, and W. H. Zurek, Phys. Rev. Lett. **77**, 198 (1996); C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Phys. Rev. A **54**, 3824 (1996).

[13] A. M. Steane, Phys. Rev. Lett. **77**, 793 (1996).

[14] B. Rahn, A. C. Doherty, and H. Mabuchi, Phys. Rev. A **66**, 032304 (2002).

[15] D. Poulin, Phys. Rev. Lett. **95**, 230504 (2005).

[16] D. Gottesman, Ph.D. thesis, California Institute of Technology, Pasadena, 1997; e-print quant-ph/9705052.

[17] I. Markov and Y. Shi, e-print quant-ph/0511069.

[18] S. Aji and R. McEliece, IEEE Trans. Inf. Theory **46**, 325 (2000).

[19] G. Forney, *Concatenated Codes* (MIT Press, Cambridge, MA, 1966).

[20] D. P. DiVincenzo, P. W. Shor, and J. A. Smolin, Phys. Rev. A **57**, 830 (1998).

[21] G. Smith and J. A. Smolin, e-print quant-ph/0604107.

[22] H. Ollivier, D. Poulin, and J.-P. Tillich (to be published).

[23] E. Knill, Nature (London) **434**, 39 (2005).

[24] A. Fletcher, P. Shor, and M. Win, e-print quant-ph/0606035.

[25] Our notion of efficient decoding differs from that of Ref. [24]: our decoding runs in linear time with the number of physical qubits.

[26] As presented here, the algorithm allows to simultaneously compute $P(L_m^{(j)})$ for all m,j. Computing only $P(L_1|\mathbf{s}_1)$ leads to simplification of the message-passing rules; in particular, messages need only to flow from the bottom to the top of the graph.