Fault-tolerant preparation of stabilizer states for quantum Calderbank-Shor-Steane codes by classical error-correcting codes

Ching-Yi Lai, Yi-Cong Zheng, and Todd A. Brun

Phys. Rev. A **95**, 032339 — Published 31 March 2017

DOI: 10.1103/PhysRevA.95.032339

Fault-tolerant Preparation of Stabilizer States for Quantum CSS Codes by Classical Error-Correcting Codes

Ching-Yi Lai,[1, *] Yi-Cong Zheng,[2, 3] and Todd A. Brun[4]

[1] *Institute of Information Science, Academia Sinica, Taipei, Taiwan 11529*
[2] *Centre for Quantum Technology, National University of Singapore, Singapore 117543*
[3] *Yale-NUS College, Singapore 138614*
[4] *Electrical Engineering Department, University of Southern California, Los Angeles, California, USA 90089.*

(Dated: March 7, 2017)

Stabilizer states are extensively studied in quantum information theory for their structures based on the Pauli group. Calderbank-Shor-Steane (CSS) stabilizer states are of particular importance in their application to fault-tolerant quantum computation (FTQC). However, how to fault-tolerantly prepare arbitrary CSS stabilizer states for general CSS stabilizer codes is still unknown, and their preparation can be highly costly in computational resources. In this paper, we show how to prepare a large class of CSS stabilizer states useful for FTQC. We propose distillation protocols using syndrome encoding by classical codes or quantum CSS codes. Along the same lines, we show that classical coding techniques can reduce the ancilla consumption in Steane syndrome extraction by using additional transversal controlled-NOT gates and classical computing power. In the scenario of a fixed ancilla consumption rate, we can increase the frequency of quantum error correction and effectively lower the error rate.

I. INTRODUCTION

Quantum states are inherently susceptible to noise, and physical devices that process quantum information are themselves generally faulty. Reliable quantum computation is still possible, however, with the help of quantum error-correcting codes. Quantum *stabilizer* codes are an especially important class of quantum codes that are similar to classical linear block codes [1], in which quantum information is encoded in the eigenstates—codewords—of a set of commuting Pauli operators called stabilizer generators.

Fault-tolerant quantum computation (FTQC) is the task of accomplishing quantum computation with arbitrary accuracy using imperfect quantum circuits. Protected by one or more stabilizer codes, a code-based FTQC scheme computes in the codespace of a stabilizer code, interspersed with repeated error corrections. A *fault-tolerant* procedure has the property that if only one component (or more generally, a small number of components) of the procedure fails, the errors produced by this failure are correctable, and are not transformed by the procedure into an uncorrectable error of the underlying error-correcting code. Threshold theorems have shown that it is possible to realize quantum computations of arbitrary size with arbitrary accuracy, provided that the errors are sufficiently local and their rates fall below a threshold [2–5]. Currently, most FTQC schemes use Calderbank-Shor-Steane (CSS) type stabilizer codes [6, 7], where every stabilizer generator can be chosen to be the tensor products of identity and either X or Z Pauli operators, and so can the logical operators. Most such FTQC schemes require the preparation of CSS stabilizer states—codewords of the CSS code that are eigenstates of some set of logical operators—for the purpose of error correction or computation.

CSS stabilizer states can be prepared by using Clifford encoding circuits (with faulty gates) [8]. However, this is not fault-tolerant, so the generated states need to be verified. Basic CSS stabilizer states, such as the logical states $|0\rangle_L$ or $|+\rangle_L$, are usually fault-tolerantly generated by specific quantum circuits in FTQC schemes [9–18]. For general CSS codes, it is not known how to produce arbitrary stabilizer states that are "clean" enough for quantum computation, especially when the code length is large.

In other contexts (e.g., entanglement purification [19, 20] or magic state distillation [21]), this problem is tackled by distillation: making a bunch of imperfect states, and then carrying out a protocol to produce a smaller number of better states. In this paper we show how *classical error-correcting codes*, together with the Steane syndrome extraction, can be applied to distill a large class of useful CSS stabilizer states (Distillation Protocol I), by actively correcting errors on a fraction of the imperfect stabilizer states that are produced by non-fault-tolerant (or fault-tolerant) methods.

If we have clean ancillas, we can use the Steane syndrome extraction to learn information about the errors—the *error syndrome* [22]. A transversal circuit is applied between the codeword and two clean ancillas, and bitwise qubit measurements are applied to the ancillas and the error syndrome is obtained by computing the parities of the corresponding measurement outcomes. However, this would obviously consume more clean ancillas than it produces. Since in our scenario only noisy stabilizer states are available, we combine Steane extraction with classical coding. After performing a transversal circuit on a set of noisy CSS stabilizer states, a subset of the states are measured bitwise and a set of parities is cal-

[*] cylai0616@iis.sinica.edu.tw

culated, and classical decoding is then applied to this set of parities to learn the error syndromes of the remaining stabilizer states. Quantum error correction can be applied accordingly to obtain clean ancillas. Along the way, we also develop a distillation protocol using quantum CSS codes rather than classical codes (Distillation Protocol II).

The only operations needed in the two protocols are *transversal* controlled-NOT (CNOT) gates, bitwise single-qubit measurements, classical decoding, and correction by applying Pauli gates (see Sec. III). These features for fault-tolerance are similar to the constraint of local operations and classical communication (LOCC) in some multipartite protocols, where each qubit is considered as a single party. Therefore our distillation protocols naturally apply to the task of multipartite entanglement purification for CSS stabilizer states by LOCC [23–28]. For simplicity, we only consider CSS codes that encode one logical qubit in this paper. Our results, however, can be generalized easily to multi-qubit codes.

The methods used for distilling ancillas—combining Steane syndrome extraction with classical error correction—can also solve a different (but related) problem. Steane syndrome extraction is used for quantum CSS codes with high-weight stabilizer generators. However, each error-correction step requires two clean ancillas per (quantum) codeword, which are of the same size as the underlying CSS codes and are expensive, especially when the code length is large. We therefore would like to use as few ancillas as possible during syndrome measurement without seriously degrading the performance of error correction. To achieve this, we propose an *ancilla saving* protocol using classical codes. Rather than using two ancillas for each code block, a smaller number of ancillas is shared among multiple code blocks, and classical decoding is used to separate out the error syndromes of the different blocks. Assuming that the error rate is low enough, this can reduce the rate of ancilla consumption for a given rate of error correction without seriously reducing its accuracy.

The paper is organized as follows. We provide preliminary material in the Sec. II, including the basics of stabilizer codes, CSS codes, and Steane syndrome extraction. The distillation protocols by classical codes and quantum CSS codes are given in Subsec. III A and Subsec. III B, respectively. Following that, we describe the ancilla saving protocol in Sec. IV. We conclude in Sec. V.

II. PRELIMINARIES

We begin with a brief review of classical codes, quantum stabilizer codes, CSS codes, and Steane syndrome extraction.

A. Classical Codes, Stabilizer Codes, and CSS Codes

Error-correcting codes protect digital information from noise by adding redundancy. The encoded information has to satisfy some mathematical relations—*parity checks*—so that errors can be detected if any of the parity checks are violated. Let H be an $(m-k)\times m$ binary matrix with full rank. Then an $[m,k,d]$ linear binary code $\mathcal{C}$ associated with parity-check matrix H is a $k-$dimensional subspace of all binary ordered $m-$tuples (row vectors) in $\mathbb{Z}_2^m$ such that

$$v\mathsf{H}^T = 0,$$

for all $v \in \mathcal{C}$, where H^T is the transpose of H and the addition is modulo 2. Such row vectors v are called *codewords* of $\mathcal{C}$. If $\mathsf{H}\tilde{v}^T \neq 0$ for some $\tilde{v} \in \mathbb{Z}_2^m$, we know that some error occurred. Hence, the rows of H are called parity-checks and $\mathsf{H}\tilde{v}^T$ is called the *error syndrome* of $\tilde{v}$. The parameter d is called the minimum distance of $\mathcal{C}$ so that any two codewords of $\mathcal{C}$ differ in at least d bits. This code can correct arbitrary $\lfloor\frac{d-1}{2}\rfloor$-bit errors. Since the code is linear, $\mathcal{C}$ can also be defined as the row space of an $k\times m$ generator matrix G, which satisfies

$$\mathsf{G}\mathsf{H}^T = 0.$$

That is, G and H are orthogonal. The dual code $\mathcal{C}^\perp$ of $\mathcal{C}$ is the row space of H. For more properties of classical codes, please refer to [29].

Now we consider the quantum case. We focus on the two-level quantum system—the *qubit*. A pure qubit state is a unit vector in the two-dimensional complex vector space $\mathbb{C}^2$ with the usual inner product and an (ordered) orthonormal basis $\{|0\rangle,|1\rangle\}$. The n-qubit state space is $\mathbb{C}^{2^n}$. Let $\mathcal{G}_n = \mathcal{G}_1^{\otimes n}$ denote the n-fold Pauli group, where

$$\mathcal{G}_1 = \{\pm I, \pm iI, \pm X, \pm iX, \pm Y, \pm iY, \pm Z, \pm iZ\}, \qquad (1)$$

and

$$I = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad X = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},$$

$$Z = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}, \quad Y = iXZ.$$

We use the notation X_j to denote $I^{\otimes j-1}\otimes X\otimes I^{\otimes n-j}$ where the underlying length n is clear from the context, and similarly for Y_j and Z_j. We also use the notation X_e for $e = e_1\cdots e_n \in \mathbb{Z}_2^n$ (row vector) to denote $\otimes_{i=1}^n X^{e_i}$, and similarly for Z_e. In general, an n-fold Pauli operator can be expressed as

$$i^{c'} \otimes_{i=1}^n X^{e_i}Z^{f_i} = i^c X_e Z_f \qquad (2)$$

for $c, c' \in \{0,1,2,3\}$ and $e, f \in \mathbb{Z}_2^n$. Thus (e,f) is called the binary representation of the Pauli operator $i^c X_e Z_f$. For example, $I \otimes X \otimes I \otimes I \otimes Z = X_2 Z_5 = X_{01000} Z_{00001}$.

The identity $I^{\otimes n}$ will be denoted by id. Let $C_i(X_j)$ denote the CNOT gate with control qubit i and target qubit j. For example, $C_1(X_2) = |0\rangle\langle 0| \otimes I^{\otimes n-1} + |1\rangle\langle 1| \otimes X \otimes I^{\otimes n-2}$. The quantum circuits in this paper consist only of CNOT gates, together with single-qubit measurements in the X or Z basis.

Suppose $\mathcal{S}$ is an Abelian subgroup of $\mathcal{G}_n$ with a set of l independent generators $\{g_1, \ldots, g_l\}$, and $\mathcal{S}$ does not include $-I$. Every element in $\mathcal{G}_n$ has eigenvalues ± 1. An $[[n, n-l]]$ quantum stabilizer code $C(\mathcal{S})$ is defined as the 2^{n-l}-dimensional subspace of the n-qubit state space ($\mathbb{C}^{2^n}$) fixed by $\mathcal{S}$, which is the joint-(+1) eigenspace of $g_1, \ldots, g_l$. Then for a codeword $|\psi\rangle \in C(\mathcal{S})$,

$$g|\psi\rangle = |\psi\rangle$$

for all $g \in \mathcal{S}$. When $l = n$, $C(\mathcal{S})$ has only one eigenstate (up to a phase) and this state is called a *stabilizer state.*

The error operators in this paper are assumed to be *Pauli errors* (i.e., operators in $\mathcal{G}_n$). This is not actually as restrictive as it sounds: since the Pauli operators form a basis, the ability to correct a set of Pauli errors implies the ability to correct a large class of general errors. If a Pauli error occurs on $|\psi\rangle$, some eigenvalues of $g_1, \ldots, g_l$ may be flipped. Therefore, we gain information about the error by measuring the stabilizer generators $g_1, \ldots, g_l$, and the measurement outcomes (in bits) of $g_1, \ldots, g_l$ are called *error syndrome.* (If the eigenvalue of a stabilizer is $+1$ or -1, its corresponding syndrome bit is 0 or 1, respectively.) Then a quantum decoder has to choose a good recovery operation based on the measured error syndromes.

CSS codes are a class of stabilizer codes whose stabilizer generators consist of the tensor products of identity and either X or Z operators [6, 7]. Let $[\mathsf{M}]_{i,j}$ denote the (i, j) entry of a matrix M. (We may also use $[v]_i$ to denote the ith entry of a vector v.) A CSS code can be defined by two matrices that are orthogonal to each other. Suppose H_Z and H_X are $r_Z \times n$ and $r_X \times n$ matrices ($r_Z + r_X \leq n$) with full rank r_Z and r_X, respectively, such that

$$\mathsf{H}_X \mathsf{H}_Z^T = 0. \tag{3}$$

Then we can define an $[[n, n - r_Z - r_X]]$ CSS code with Z stabilizer generators

$$g_i = \bigotimes_{j=1}^{n} Z^{[\mathsf{H}_Z]_{i,j}}, \quad i = 1, \ldots, r_Z, \tag{4}$$

and X stabilizer generators

$$g_{r_Z+i} = \bigotimes_{j=1}^{n} X^{[\mathsf{H}_X]_{i,j}}, \quad i = 1, \ldots, r_X. \tag{5}$$

The condition in Eq. (3) implies that the X and Z stabilizer generators all commute.

The *check matrix* of an $[[n, n-l]]$ stabilizer code is an $l \times 2n$ binary matrix whose rows are the binary representations (Eq. (2)) of the stabilizer generators $g_1, \ldots, g_l$. For the CSS code defined by H_Z and H_X, its check matrix is

$$\begin{bmatrix} \mathsf{H}_Z & 0 \\ 0 & \mathsf{H}_X \end{bmatrix}. \tag{6}$$

The error syndrome of a Pauli error is the string of binary outcomes of measuring the stabilizers $g_1, \ldots, g_l$. Since a Pauli error can be expressed as a product of X and Z operators, error correction can be done by treating Pauli X and Z errors separately. For CSS codes, the eigenvalues of $g_1, \ldots, g_{r_Z}$ (respectively, $g_{r_Z+1}, \ldots, g_l$) correspond to the error syndrome of X (resp. Z) errors. Suppose a Pauli error $X_e Z_f$ occurs on a codeword, where $e, f \in \mathbb{Z}_2^n$. Then its (binary) X error syndrome $s_X \in \mathbb{Z}_2^{r_Z}$, which corresponds to the eigenvalues of $g_1, \ldots, g_{r_Z}$, is given by

$$s_X = e\mathsf{H}_Z^T, \tag{7}$$

and the Z error syndrome is defined similarly:

$$s_Z = e\mathsf{H}_Z^T. \tag{8}$$

In addition to the stabilizer elements, there will in general also be Pauli operators that commute with all the stabilizer generators without being in the stabilizer themselves. We call these *logical operators.* For a CSS code, it is always possible to find a subset of logical operators involving only the identity and either X or Z operators, which generate all other logical operators up to multiplication by a stabilizer element. Moreover, these logical generators can be chosen in anticommuting pairs, usually denoted $\bar{X}_j$, $\bar{Z}_j$, where each pair corresponds to logical qubit j in the code. For simplicity, in the rest of this paper we consider the preparation of stabilizer states of CSS codes $\mathcal{Q}$ that encode one logical qubit (or $r_Z + r_X = n - 1$). Our results can be directly generalized to the case of $[[n, n - r_Z - r_X]]$ multiple-qubit CSS codes ($n - r_Z - r_X > 1$). Let $\bar{X}$ and $\bar{Z}$ denote the logical X and Z operators of $\mathcal{Q}$. A quantum state with L in the subscript refers to an encoded state. For example, the encoded $|0\rangle$, $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ are denoted by $|0\rangle_L$ and $|+\rangle_L$, respectively.

B. CSS Stabilizer States and Steane Syndrome Extraction

Steane suggested a method to extract error syndromes for CSS codes [22], as shown in Fig. 1. Two clean ancillas $|+\rangle_L$ and $|0\rangle_L$ in the logical states of the underlying CSS code $\mathcal{Q}$ are used to measure the X and Z error syndromes, respectively. Each CNOT gate in Fig. 1 represents transversal CNOT gates, and the measurements

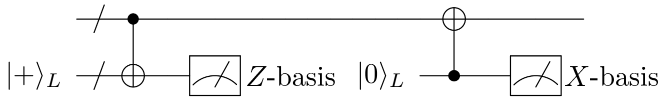

FIG. 1. Quantum circuit for Steane syndrome extraction.

in the Z or X basis are bitwise. In the circuit, X and Z errors on the data qubit will propagate, respectively, to the ancillas $|+\rangle_L$ and $|0\rangle_L$ through the CNOTs, so that we learn error information by measuring the two ancillas.

Suppose the measurement outcomes of $|+\rangle_L$ and $|0\rangle_L$ are m_X and m_Z (in bits), respectively. Then the (observed) X and Z syndromes are computed by $m_X\mathsf{H}_Z^T$, and $m_Z\mathsf{H}_X^T$, respectively. We can perform error correction according to these syndromes or just keep track of them.

The two ancillas $|+\rangle_L$ and $|0\rangle_L$ are actually stabilizer states of $\mathcal{Q}$ by including logical operator $\bar{X}$ or $\bar{Z}$ in with the stabilizer generators. That is, $|+\rangle_L$ is stabilized by $\langle g_1,\dots,g_{n-1},\bar{X}\rangle$, and $|0\rangle_L$ is stabilized by $\langle g_1,\dots,g_{n-1},\bar{Z}\rangle$. These two stabilizer states can be produced by Clifford encoding circuits with CNOT gates only, together with the ability to prepare physical qubits in $|0\rangle$ and $|+\rangle$ [30]. If imperfect quantum gates are used in the encoding circuit, the states must be verified.

We can use Steane syndrome extraction to correct errors on a noisy ancilla, but it requires us already to have two clean ancillas, which is clearly impractical. In the following section, we will introduce a protocol for distilling a number of CSS stabilizer states from a larger set of imperfect ones by using classical error-correcting codes and quantum CSS codes, and demonstrate how to distill the states $|0\rangle_L$ or $|+\rangle_L$.

A simple example of a CSS stabilizer state is the Einstein-Podolsky-Rosen (EPR) pair:

$$\frac{|00\rangle+|11\rangle}{\sqrt{2}},$$

which is stabilized by $X\otimes X$ and $Z\otimes Z$. Córcoles *et al.* recently experimentally demonstrated error correction on an EPR pair [31]. EPR pairs were the first states for which a distillation protocol was proposed [19, 20].

III. ANCILLA DISTILLATION

Suppose we are using an $[[n,1]]$ CSS code $\mathcal{Q}$ defined by matrices $\mathsf{H}_Z,\mathsf{H}_X$ with $r_Z+r_X=n-1$ as in the previous section. Given a noisy stabilizer state, e.g., $X_e|0\rangle_L$, we can perform quantum error correction and restore the state up to some logical operator if we know the actual error syndrome. Now, measuring the logical operator $\bar{Z}$ will tell us whether there is a logical error or not, and additional logical correction can be applied if necessary. Thus we can, ideally, have a *perfect* stabilizer state $|0\rangle_L$. In reality we are in a situation where the ancillas for syndrome measurements are also imperfect, and we will address this issue in this section.

Suppose we are given a bunch of imperfect ancillas in some CSS stabilizer state, e.g., $|0\rangle_L$ or $|+\rangle_L$, and we wish to purify them. Our approach is to determine the correct error syndromes of a subset of the ancillas by measuring the rest. More precisely, we will use a transversal quantum circuit to couple m noisy ancillas according to a classical error-correcting code, measure $m-k$ of them, and then extract the error syndromes of the remaining k ancillas. This procedure is called *ancilla distillation by classical codes* and will be detailed in the first subsection. In the second subsection, we generalize the idea to distill the noisy ancillas by using an arbitrary quantum CSS code.

For now, we neglect any errors in the CNOTs or measurements used in the quantum circuit for distillation. We discuss the issue of noisy distillation circuits in Section V.

A. Distillation by Classical Codes

The key observation is that classical binary linear codes can be encoded or decoded by circuits using only CNOTs. Suppose $\mathcal{C}_D$ is an $[m,k,d]$ binary linear block code that can correct $t=\lfloor\frac{d-1}{2}\rfloor$ errors. Such a code has $r=m-k$ parity checks. Let $\mathsf{H}_D=[\mathsf{A}^T\ \mathsf{I}_r]$ be the parity-check matrix of $\mathcal{C}_D$ in systematic form, where I_k is the $k\times k$ identity matrix and A is $k\times r$. We define a quantum distillation circuit U_D by

$$U_D=\prod_{i=1}^{k}\prod_{j=1}^{r}C_i(X_{k+j})^{[\mathsf{A}]_{i,j}}. \tag{9}$$

That is, $C_i(X_{k+j})$ is applied if $[\mathsf{A}]_{i,j}=1$, and id is applied, otherwise. Consider $X_e|0^m\rangle=|e\rangle$, where $e=e_1\cdots e_m\in\mathbb{Z}_2^m$ and

$$0^m=\underbrace{0\cdots0}_{m}.$$

Then

$$U_D|e\rangle=|e_1\cdots e_k\rangle\otimes|s_e\rangle,$$

where $s_e^T=\mathsf{H}_De^T$ is the classical error syndrome of e with respect to $\mathcal{C}_D$. Then we can use a decoder of $\mathcal{C}_D$ to find the most probable error vector $\tilde{e}\in\mathbb{Z}_2^m$ and then correct the bit-flip errors in $|e_1\cdots e_k\rangle$. This decoding procedure is the main conceptual tool of our distillation protocol.

Distillation of CSS stabilizer states by classical codes involves two rounds of error correction: one for X errors and one for Z errors.

Distillation Protocol I:

1) Using an encoding circuit, we prepare many noisy copies of an n-qubit CSS stabilizer state. Divide the noisy ancillas up into groups of m.

2) (Round 1: X errors) In each group of m ancillas, choose the last r of the ancillas to hold the parity checks, and apply U_D *transversally*: that is, apply U_D to the first qubits of all m ancillas in the group, to the second qubits, and so forth. This unitary U_D applies transversal CNOTs according to the pattern of 1s in the binary matrix A.

3) Measure every qubit in the r parity-check ancillas (the last r ancillas in each group of m) in the Z basis. Let the binary row vectors $\nu^{(1)},\dots,\nu^{(r)} \in \mathbb{Z}_2^n$ be the outcomes of these measurements.

4) Calculate $\sigma^{(i)} \triangleq \nu^{(i)}\mathsf{H}_Z^T$ for $i = 1,\dots,r$. For $j = 1,\dots,r_Z$, use $[\sigma^{(1)}]_j\cdots[\sigma^{(r)}]_j$ as a classical error syndrome of $\mathcal{C}_D$ and use a decoder of $\mathcal{C}_D$ to find the most probable error vector $\tilde{s}_j^{(1)}\cdots\tilde{s}_j^{(m)}$ with this error syndrome. Then $\tilde{s}_1^{(i)}\cdots\tilde{s}_{r_Z}^{(i)}$ is the estimated X error syndrome of the ith target ancilla. Correct the X errors, if any (or just keep track of them).

5) If the distillation target is $|0\rangle_L$ or $|1\rangle_L$, calculate the parity of $\bar{Z}$ from $\nu^{(i)}$ and estimate the syndrome bits for the logical operators as in the previous step. Correct the logical errors $\bar{X}$, if any (or just keep track).

6) Of our original large number of ancillas, a fraction k/m are left. Again, divide them up into groups of m. It is very important that ancillas that were grouped together in the first round are not grouped together in the second round, because their errors are correlated.

7) (Round 2: Z errors) Similarly to step 2 above, do a transversal U_D^H, where

$$U_D^H = \prod_{i=1}^{k}\prod_{j=1}^{r} C_{k+j}(X_i)^{[\mathsf{A}]_{i,j}}. \tag{10}$$

(The control and target qubits of U_D are switched to obtain U_D^H.)

8) Measure the r parity-check ancillas in the X basis. Then repeat steps 3–5, but with X and Z switched everywhere, and with $|0\rangle,|1\rangle$ replaced by $|+\rangle,|-\rangle$, respectively. □

This procedure is somewhat technical, so we will demonstrate the distillation protocol with a detailed example.

Example 1. Fig. 2 illustrates the quantum circuit for distillation by the $[3,1,3]$ repetition code with a parity-check matrix

$$H_D = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \tag{11}$$

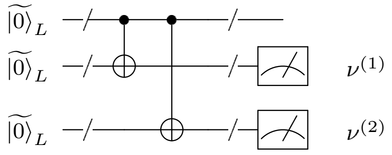
$\widetilde{|0\rangle}_L$
$\widetilde{|0\rangle}_L$
$\widetilde{|0\rangle}_L$
$\nu^{(1)}$
$\nu^{(2)}$

FIG. 2. The circuit for distilling $|0\rangle_L$ by the classical $[3,1,3]$ repetition code, where $\widetilde{|0\rangle}_L$ are imperfect logical states. The last two $\widetilde{|0\rangle}_L$ serve as parity-check ancillas.

for the ancilla state $|0\rangle_L$ of an n-qubit quantum CSS code. We now demonstrate the above protocol by distilling the ancillas of the [[7, 1, 3]] Steane code [32], with stabilizer generators

$$\begin{aligned} g_1 &= Z_1Z_4Z_5Z_7,\\ g_2 &= Z_2Z_4Z_6Z_7,\\ g_3 &= Z_3Z_5Z_6Z_7,\\ g_4 &= X_1X_4X_5X_7,\\ g_5 &= X_2X_4X_6X_7,\\ g_6 &= X_3X_5X_6X_7, \end{aligned}$$

and logical operators $\bar{X} = X_1X_2X_4$, $\bar{Z} = Z_1Z_2Z_4$.

Suppose the $[3,1,3]$ repetition code is used for distilling several noisy codewords of the Steane code. First, the noisy codewords are divided in to groups of $m = 3$. Consider one group with the three noisy Steane codewords $E_1|0\rangle_L$, $E_2|0\rangle_L$, and $E_3|0\rangle_L$ prepared independently with errors $E_{1,2,3}$, where $E_1 = X_1X_2X_4 = \bar{X}$, $E_2 = X_3$, and $E_3 = X_6X_7$. Note that E_1 and E_3 are uncorrectable errors for the Steane code. The X error syndromes are

$$\begin{array}{r l} E_1 : & 000 \,|\, 1\\ E_2 : & 001 \,|\, 0\\ E_3 : & 110 \,|\, 0 \end{array}$$

with respect to g_1, g_2, g_3, and $\bar{Z}$, respectively. After the (perfect) distillation circuit U_D by the $[3,1,3]$ code (Fig. 2), the errors become $E_1' = E_1$, $E_2' = E_1E_2 = X_1X_2X_3X_4$, and $E_3' = E_1E_3 = X_1X_2X_4X_5X_6$. Then measuring bitwise the second and the third codewords, and calculating the parities of g_1, g_2, g_3 and $\bar{Z}$, we have their syndrome bits

$$\begin{array}{r l} \sigma^{(1)} : & 001 \,|\, 1,\\ \sigma^{(2)} : & 110 \,|\, 1. \end{array}$$

Now we can use the parity check matrix of the $[3,1,3]$ repetition code to recover the four syndrome bits of the first codeword:

$$\tilde{s}^{(1)} :\ 000 \,|\, 1.$$

Since the fourth bit is 1, we apply logical operator $\bar{X}$ to the first codeword to correct the logical error and the final state is $|0\rangle_L$. Thus we have fault-tolerantly prepared an ancilla $|0\rangle_L$. □

Now we carry out protocol I for general CSS codes as follows. Suppose the X errors on the m ancillas are $X_{e^{(1)}}$, …, $X_{e^{(m)}}$, where $e^{(j)} = e_1^{(j)} \cdots e_n^{(j)} \in \mathbb{Z}_2^n$. Let $s^{(1)} = e^{(1)}\mathsf{H}_Z^T, \ldots, s^{(m)} = e^{(m)}\mathsf{H}_Z^T$. The decoding operator U_D will transform $X_{e_j^{(1)}} \otimes X_{e_j^{(2)}} \otimes \cdots \otimes X_{e_j^{(m)}}$ to

$$X_{e_j^{(1)}} \otimes \cdots \otimes X_{e_j^{(k)}} \otimes X_{s_X^j}$$

for $j = 1, \ldots, n$, where $s_X^j \in \mathbb{Z}_2^r$ is the error syndrome of $e_j^{(1)} \cdots e_j^{(m)} \in \mathbb{Z}_2^m$ with respect to H_D. If we could measure s_X^j, we could decode e_j^i directly and then obtain k clean ancillas.

However, the measurement outcomes $\nu^{(1)}, \ldots, \nu^{(r)}$ of step 3 are actually a disturbed version of s_X^j:

$$\begin{aligned}\left(\nu^{(1)^T} \cdots \nu^{(r)^T}\right) &= \begin{pmatrix} s_X^1 \\ \vdots \\ s_X^n \end{pmatrix} + \left(c_1^T \cdots c_r^T\right), \\ &= \left(e^{(1)^T} \cdots e^{(m)^T}\right) \mathsf{H}_D^T + \left(c_1^T \cdots c_r^T\right),\end{aligned} \tag{12}$$

where the row vectors $\{c_j\}$ are unknown codewords of the classical code with parity check matrix H_Z. Since we do not know the codewords, we cannot learn s_X^j.

On the other hand, we still *can* learn the quantum error syndrome

$$s^{(1)} = e^{(1)}\mathsf{H}_Z^T, \ldots, s^{(m)} = e^{(m)}\mathsf{H}_Z^T$$

from $\nu^{(1)}, \ldots, \nu^{(r)}$. Multiplying (12) from the right by H_Z^T, we have

$$\mathsf{H}_D \begin{bmatrix} s^{(1)} \\ \vdots \\ s^{(m)} \end{bmatrix} = \begin{bmatrix} \nu^{(1)}\mathsf{H}_Z^T \\ \vdots \\ \nu^{(r)}\mathsf{H}_Z^T \end{bmatrix} \triangleq \begin{bmatrix} \sigma^{(1)} \\ \vdots \\ \sigma^{(r)} \end{bmatrix}. \tag{13}$$

Then we can choose any decoder of $\mathcal{C}_D$ to recover the k error syndromes $\tilde{s}^{(i)}$ of the target ancillas as in step 4. That is, we are retrieving a particular syndrome bit j of every ancilla $\tilde{s}_j^{(1)}, \ldots, \tilde{s}_j^{(m)}$ in one classical decoding procedure. From that, we can correct all the X errors in the target k ancillas (and also any logical X errors if we are distilling $|0\rangle_L$).

Remark 1. As long as fewer than $\lfloor \frac{d-1}{2} \rfloor$ of the m syndrome bits $s_j^{(1)}, \ldots, s_j^{(m)}$ are 1s, they can be recovered by the classical decoding, assuming that the quantum gates in the distillation circuit are perfect.

Remark 2. The distillation protocol depends only on the error-correcting power of the classical code $\mathcal{C}_D$, and not on the error-correcting ability of the stabilizer states being distilled. If $n = 1$, this procedure reduces to the standard classical decoding of $\mathcal{C}_D$.

After round 1, the remaining k ancillas from the group will have lower rates of X errors than they started with. However, Z errors on the parity-check ancillas can propagate via the CNOTs back onto these k ancillas, increasing the rate of Z errors (and also correlating the errors across the ancillas). How do these changes compare? Assume that the original rates of Z and X errors are both p. The rate of Z errors on the remaining k ancillas will increase to $\sim (\beta+1)p$, where $\beta \leq r$ is the number of parity checks that each qubit is included in (or the number of 1s in each row of A). The rate of X errors goes from p to cp^{t+1}, where c is a constant that depends on the details of the codes. If p is not too big, the rate of Z errors has grown roughly by a constant factor, while the rate of X errors has been substantially reduced.

After round 2, we are left with a fraction $(\frac{k}{m})^2$ of our original ancillas. The rate of Z errors will go from $(\beta+1)p$ to $c((\beta+1)p)^{t+1} = c'p^{t+1}$, and the rate of X errors will go from cp^{t+1} to $(\beta+1)cp^{t+1} = c''p^{t+1}$. (So the rate of an arbitrary Pauli error is roughly $\tilde{c}p^{t+1}$ for some $\tilde{c}$.) These constants will not generally be equal to each other (and indeed, the starting rates for X and Z might not have been equal); one might use two different classical error-correcting codes in the two rounds so that the final rates both end up below some desired fraction. To reach a desired target error rate, this procedure could be iterated; or one could just vary the distances of the classical codes used depending on the original error rates and the desired final error rates.

Example 2. In addition to the $[3,1,3]$ repetition code above, we simulated distillation by the $[5,1,5]$ repetition code and the $[7,4,3]$ Hamming code [29]. The results are shown in Fig. 3. The simulations begin with preparation of noisy ancillas by the CSS encoding circuit. We assume that during the ancilla preparation each individual qubit suffers independent depolarizing errors with rate p. Noisy CNOTs in the encoding circuit are modeled as a perfect CNOT, followed by no error (with probability $1-p$) or one of the nonidentity two-fold Paul operators (e.g., $X \otimes Y$, $I \otimes Z$, $X \otimes I$, …) with equal probabilities $p/15$. For our initial analysis, we assume that the distillation circuit does not itself contain errors; the issue of imperfect distillation circuits will be addressed in the discussion section.

We define the final error rate to be the probability of any Pauli error left in the target ancillas after *two rounds* of distillation. This estimate is *pessimistic*, since it ignores the likelihood that the residual errors are correctable in the next error correction cycle.

After the noisy encoding circuit, the error rate on each qubit will increase to $\sim (\alpha+1)p$, where α is the number of stabilizer generators each qubit is involved. If p is small enough, the distillation protocol will work. As can be seen in the logarithmic plot in Fig. 3, each curve appears linear with slope $t+1$ and the "threshold" for each code is specified by $\log p_{\text{th}} = -\frac{1}{t}\log\tilde{c}$. (By "threshold," we mean the crossing point of a curve with the dashed line. That is, the point where the distillation error rate and the physical rate p are equal.) Thus, we say that a stabilizer state can be fault-tolerantly prepared if the error rate is below the threshold for the classical code used in distillation. □

For these simple examples, we can see that the [5,1,5] code with higher distance has asymptotically better behavior in the low physical error rate regime but the yield rate is low. On the other hand, the [7,4,3] code has a larger yield rate but worse performance. Since there are many efficient classical codes with high rate and good error-correcting ability, we can certainly find better candidates for the protocol with extra cost.

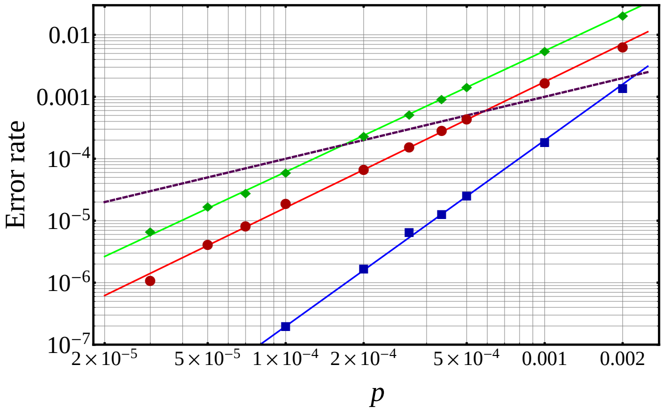

FIG. 3. (Color online.) Ancilla distillation by 1) the $[7,4,3]$ Hamming code (green diamonds); 2) the $[3,1,3]$ repetition code (red dots); 3) the $[5,1,5]$ repetition code (blue squares). The dashed line is the rate without distillation. Up to 7×10^8 iterations are used for each point.

Finally we briefly discuss candidates for large quantum codes whose raw ancilla state preparation by noisy Clifford circuits will have sufficiently low error rates. It is known that quantum CSS codes built from classical doubly-even codes allow transversal Clifford gates [33, 34]: such codes as the Steane code, the $[[23,1,7]]$ quantum Golay code, or other quantum quadratic-residue codes [35]. We can concatenate a large quantum block code with, for example, the quantum Golay code, whose Clifford operations can be done transversally, and these ancillas can be prepared fault-tolerantly [12]. With quantum Golay code blocks at the bottom level, the Clifford encoding circuit for the large quantum code can be transversally implemented, and quantum error correction by the quantum Golay code can be inserted into the circuit, if necessary. Therefore, the error rate of the raw ancillas can be suppressed to below the "threshold" of distillation. If we constantly apply quantum error correction by the Golay code, the output should be very good, but the overhead will be large.

Another intriguing possibility is to perform distillation steps at several points during the encoding circuit for the ancillas, to remove errors before they can spread widely. Efficient distillation would require the use of very high-rate classical codes.

B. Distillation by Quantum CSS Codes

Instead of using classical codes in two steps to correct both X and Z errors, we can similarly use an $[[m,k]]$ quantum CSS code $\mathcal{Q}_D$ to distill the desired ancillas of an $[[n,1]]$ code. For simplicity, suppose $\mathcal{Q}_D$ is defined by H'_Z and H'_X of dimension $r'_Z\times m$ and $r'_X\times m$ matrices $(k=m-(r'_Z+r'_X)>0)$ with full rank r'_Z and r'_X and $\mathsf{H}'_X\mathsf{H}'^T_Z=0$. $\mathcal{Q}_D$ has r'_Z Z generators, and r'_X X generators. The check matrix of an $[[m,k]]$ CSS code can be written in the following form [8]:

$$\begin{pmatrix} \mathsf{H}'_Z & 0 \\ 0 & \mathsf{H}'_X \end{pmatrix} = \left(\begin{array}{ccc|ccc} \mathsf{I}_{r'_Z} & \mathsf{A} & \mathsf{B} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{D} & \mathsf{I}_{r'_X} & \mathsf{F} \end{array}\right),$$

where A, B, D, and F are binary matrices of appropriate dimensions. We use a particular encoding circuit as follows: to the k information qubit state $|\phi\rangle$, append r'_Z ancilla qubits in the state $|0\rangle$ and r'_X ancilla qubits in the state $|+\rangle$, which will correspond to the r'_Z Z generators and r'_X X generators, respectively, after encoding. That is, we apply an encoding circuit to the initial state

$$|0\rangle^{\otimes r'_Z}\otimes|+\rangle^{\otimes r'_X}\otimes|\phi\rangle,$$

which corresponds to a check matrix

$$\left(\begin{array}{ccc|ccc} \mathsf{I}_{r'_Z} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathsf{I}_{r'_X} & 0 \end{array}\right).$$

The encoding circuit U'_D, which is a unitary operator, will then consist only of CNOT gates [30]. It can be verified that this works for any CSS codes. and we postpone this justification to Appendix A.

Suppose an error operator $E\in\mathcal{G}_m$ occurs on a codeword of $\mathcal{Q}_D$. We decode by running the above encoding circuit backwards, so that the error syndrome information will be contained in the ancilla qubits. After the decoding circuit ${U'_D}^\dagger$, we have the following transformed error operating on the initial state $|0\rangle^{\otimes r'_Z}\otimes|+\rangle^{\otimes r'_X}\otimes|\phi\rangle$:

$${U'_D}^\dagger E U'_D \triangleq X_{s_X}\otimes Z_{s_Z}\otimes L_E,$$

where $s_X\in\mathbb{Z}_2^{r'_Z}$, $s_Z\in\mathbb{Z}_2^{r'_X}$ are the syndrome vectors (hence $X_{s_X}\in\mathcal{G}_{r'_Z}$, $Z_{s_Z}\in\mathcal{G}_{r'_X}$) and $L_E\in\mathcal{G}_k$ is an logical error operator.

We then measure the first r'_Z ancillas in the Z basis, which gives the X error syndrome s_X, and the other r'_X ancillas in the X basis, which gives the Z error syndrome s_Z. From that, we can figure out what corrections, if any, need to be applied to the k information qubits. This leads to the following distillation scheme:

Distillation Protocol II:

Suppose we want to distill a CSS stabilizer state of the

[[$n,1$]] code $\mathcal{Q}$ defined at the beginning of this section.

1) Start with m imperfect copies of a stabilizer state of $\mathcal{Q}$. Do a transversal ${U'_D}^\dagger$ on the m copies of the n-qubit system.

2) For the r'_Z (respectively r'_X) systems in the positions corresponding to the Z (resp. X) ancillas, we measure all the qubits in the Z (resp. X) basis and let $\nu^{(1)},\dots,\nu^{(r'_Z)}\in\mathbb{Z}_2^m$ (resp. $\nu^{(1)},\dots,\nu^{(r'_X)}\in\mathbb{Z}_2^m$) be the outcomes.

3) Calculate $\sigma_X^{(i)}\triangleq\nu_Z^{(i)}\mathsf{H}_Z^T$ for $i=1,\dots,r'_Z$. For $j=1,\dots,r_Z$, use $[\sigma_X^{(1)}]_j\cdots[\sigma_X^{(r'_Z)}]_j$ as a classical error syndrome with respect to the parity-check matrix H'_Z and use a corresponding decoder to find the most probable error vector $\tilde{s}_j^{(1)}\cdots\tilde{s}_j^{(m)}$ with this error syndrome. Then $\tilde{s}_1^{(i)}\cdots\tilde{s}_{r_Z}^{(i)}$ is the estimated X error syndrome of the ith target ancilla. We can correct the X errors, if any, (or just keep track of them).

4) Repeat 3) but with X and Z switched everywhere.

5) If the distillation target is $|0\rangle_L$ or $|1\rangle_L$ (resp. $|+\rangle_L$ or $|-\rangle_L$), then calculate the parity of $\bar{Z}$ (resp. $\bar{X}$) from $\nu_Z^{(i)}$ (resp. $\nu_X^{(i)}$) and estimate the syndrome bits for the logical operators as in the previous step. Correct the logical error $\bar{X}$ (resp. $\bar{Z}$), if any, (or just keep track of them).

□

This protocol is very similar to Distillation Protocol I. Thus we omit the explanation of how it works. We end up with k much cleaner copies of the n-qubit CSS stabilizer states of $\mathcal{Q}$. Again, this procedure could be iterated, or we can use a good enough [[m,k]] code. This is like a concatenation of the n-qubit code $\mathcal{Q}$ with the [[m,k]] code $\mathcal{Q}_D$, but only the decoding on the second level, $\mathcal{Q}_D$, is applied.

Protocol I by classical codes is more flexible than Protocol II, since any classical codes can be used in Protocol I and the codes can be different in two rounds of distillation; whereas only dual-containing codes can be applied in Protocol II. On the other hand, dual-containing codes used in Protocol II can also be applied in Protocol I, and the number of CNOTs for correcting both X and Z errors are roughly the same in the two protocols. Basically, the performance of these two protocols are strongly related to the performance of the classical codes, so we omit simulations of Protocol II here.

IV. ANCILLA SAVING

Clean ancillas $|0\rangle_L$ and $|+\rangle_L$ in Steane syndrome extraction are expensive resources. We would like to use as few of them as possible during syndrome measurement, as long as errors do not accumulate seriously. In the following, we show that classical coding can also reduce ancilla consumption in Steane syndrome extraction; it turns out that this problem is equivalent to the distillation problem in Sec. III A. We assume that the quantum circuits for error correction are perfect, and the ancillas are assumed to be *clean* in the following discussion.

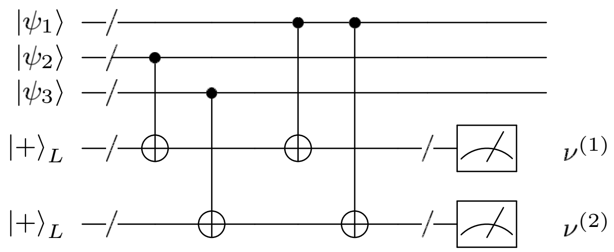

FIG. 4. The circuit for syndrome measurement of three data blocks with two ancilla blocks by the $[3,1,3]$ repetition code.

Suppose we have m codewords $|\psi_1\rangle,\dots,|\psi_m\rangle$ of the [[$n,1$]] CSS code $\mathcal{Q}$ defined by H_Z and H_X. Let $X_{e^{(j)}}Z_{f^{(j)}}$ be the error corrupting $|\psi_j\rangle$ for $j=1,\dots,m$. If Steane syndrome extraction is used, m (perfect) ancillas $|+\rangle_L$ are required to measure the X error syndromes $e^{(j)}\mathsf{H}_Z^T$. Our goal here is to estimate the X error syndromes by using only r ($<m$) clean ancillas $|+\rangle_L$. The treatment for Z error is similar.

Let $\mathcal{C}_D$ be an $[m,k=m-r,d]$ classical code with parity-check matrix $\mathsf{H}_D=[\mathsf{A}^T\ \mathsf{I}_r]$. Assume we have r clean ancillas $|+\rangle_L$ and r clean ancillas $|0\rangle_L$.

Ancilla Saving Protocol:

1) Apply transversal CNOTs from $|\psi_j\rangle$ to the j-th $|+\rangle_L$ for $j=k+1,\dots,m$.

2) Apply a transversal CNOT from $|\psi_i\rangle$ to the j-th $|+\rangle_L$ if $[\mathsf{A}]_{i,j}=1$.

3) Do steps 3 and 4 of Distillation Protocol I.

4) Apply transversal CNOTs from the j-th $|0\rangle_L$ to $|\psi_j\rangle$ for $j=k+1,\dots,m$.

5) Apply a transversal CNOT from the j-th $|0\rangle_L$ to $|\psi_i\rangle$ to if $[\mathsf{A}]_{i,j}=1$.

6) Do steps 3 and 4 of Distillation Protocol I, but with X and Z switched everywhere.

□

Fig. 4 demonstrates the circuit for the X syndrome extraction of three data blocks with two ancilla blocks by the $[3,1,3]$ repetition code. Observe that this circuit is essentially equivalent to the circuit in Fig. 2. We can combine the X errors from the second and third encoded states with the two clean ancillas $|+\rangle_L$, respectively, and then remove those two encoded states. As a consequence, our ancilla saving protocol is equivalent to the distillation protocol by classical codes.

We may compare the error correction performance of this ancilla saving protocol by $\mathcal{C}_D$ on codewords of $\mathcal{Q}$, say

$\mathcal{Q}+\mathcal{C}_D$, with the original Steane extraction scheme. A good figure of merit for comparison is the channel fidelity of a quantum code over a noise channel [36, 37]. For simplicity, here we assume the independent single-qubit noise channel is

$$\mathcal{E}(\rho)=(1-p)\rho+pX\rho X,$$

for any single-qubit state ρ. Unquestionably the channel fidelity is expected to drop with fewer ancillas in the saving protocol, and both the encoding and decoding complexities will increase. Thus, we have a tradeoff between channel fidelity, gate complexity, and ancilla consumption.

Example 3. Consider the $[[7,1,3]]$ Steane code. It is known that the channel fidelity of a quantum code over a Pauli channel is the probability of correctable errors [38]: that is, the probability of a set of coset leaders and their degenerate errors. It is more complicated to calculate the channel fidelity of the saving protocol. We estimate it as follows:

Let $F_C(\mathcal{E})_i$ be the channel fidelity of sending $|\psi_i\rangle$ through $\mathcal{E}^{\otimes n}$ for $i=1,\ldots,m$. Then the average channel fidelity is

$$\overline{F_C}(\mathcal{E})=\frac{1}{m}\sum_i F_C(\mathcal{E})_i.$$

Since $|\psi_i\rangle$ are correlated in the saving protocol, the previous result of $F_C(\mathcal{E})$ cannot be applied here. Thus we apply Monte Carlo methods to approximate $\overline{F_C}(\mathcal{E})$:

1) Fix p. Set $i:=1$.

2) Apply Pauli X errors $X^i_{e(1)},\ldots,X^i_{e(m)}$ to perfect information states $|\psi_1\rangle,\ldots,|\psi_m\rangle$, where $X^i_{e(j)}\in\mathcal{P}_n$ is an n-fold Pauli X error, randomly generated according to the probability distribution of $\mathcal{E}$.

3) Use the ancilla-saving code $\mathcal{C}$ to recover the error syndromes $s_1,\ldots,s_m$ for $X^i_{e(1)},\ldots,X^i_{e(m)}$, respectively. If there is no logical error on $|\psi_j\rangle$ after decoding, $X^i_{e(j)}$ is correctable. Set $i:=i+1$.

4) Repeat steps 2 and 3 N times.

5) Count the number of correctable errors in $\{X^i_{e(j)}\}$, say M. Then $\overline{F_C}(\mathcal{E})$ is approximated by $\frac{M}{mN}$.

Fig. 5 illustrates plots on $\overline{F_C}(\mathcal{E})$ for various ancilla saving protocols: the $[7,4,3]$ Hamming code, and the $[3,1,3]$ and $[5,1,5]$ repetition codes, together with $F_C(\mathcal{E})$ without ancilla-saving. As shown in Fig. 5, at $p=0.01$, the channel fidelity for $[[7,1,3]]+[3,1,3]$ drops from 0.998 to 0.988. Since only two ancilla states are used, we reduce the ancilla consumption by 33.3% by using one additional transversal CNOT and increased classical computing complexity. (Note that the circuit for preparing a $|+\rangle_L$ has nine CNOTs.) For the $[7,4,3]$ code, the fidelity drops significantly to save $\frac{4}{7}=57.1\%$ ancillas. For the $[5,1,5]$ code, the fidelity drop at $p=0.01$ is less than 0.2%, while $\frac{1}{5}=20\%$ of the ancillas are saved.

□

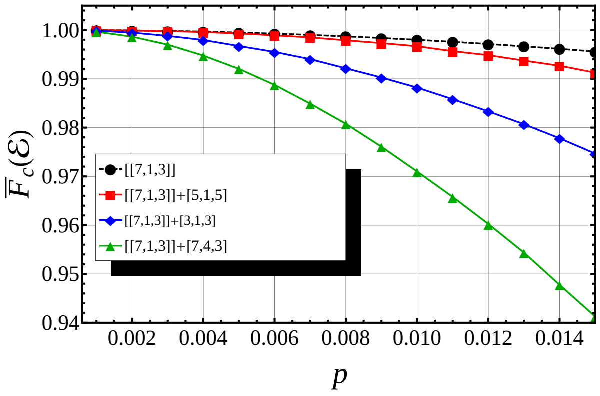

FIG. 5. (Color online.) The average channel fidelity of the Steane code without or with ancilla saving by the $[3,1,3]$, $[7,4,3]$, or $[5,1,5]$ code. The number of iterations for each point is up to 7×10^8.

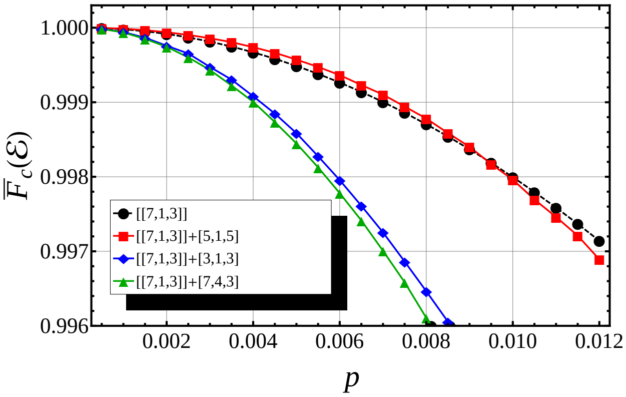

FIG. 6. (Color online.) The average channel fidelity of Steane code with ancilla saving by the $[3,1,3]$, $[7,4,3]$, or $[5,1,5]$ code at physical error rate p. The number of iterations for each point is up to 7×10^8.

When the ancilla consumption rate is fixed, we can increase the frequency of quantum error correction with the ancilla saving protocol, which is equivalent to lowering the error rate on the data qubits. Let us define the effective error rate as the (accumulated) physical error rate between two error correction steps. Here the effective error rate of an $[[n,1]]$ code without ancilla saving is simply called the *physical error rate.* If the physical error rate is p, then the effective error rate of an $[[n,1]]+[m,m-r]$ ancilla saving protocol is rp/m, assuming that quantum error correction is sufficiently fast. Let F^p_{o} and F^p_{comb} be the channel fidelities of the original and the $[[n,1]]+[m,m-r]$ protocols at *effective* error rate p, respectively. Apparently we have $F^0_{\text{o}}=F^0_{\text{comb}}=1$ and $F^p_{\text{o}}>F^p_{\text{comb}}$ for $p>0$. Thus by the continuity of fidelity, there exists p^*

such that $F_{\text{o}}^{p^*} = F_{\text{comb}}^{rp^*/m}$. Here we define the *effective channel fidelity* of an $[[n,1]]+[m,m-r]$ protocol as the average channel fidelity of the protocol at effective error rate rp/m. Hence for physical error rate $p < p^*$ the effective channel fidelity of the $[[n,1]]+[m,m-r]$ protocol is higher. Let us consider the above example again. Fig. 6 plots these channel fidelities. As can be seen, applying the ancilla saving protocol with the $[5,1,5]$ code is better than the original scheme for $p < 0.00925$, but there is no fidelity gain for the other two codes. Of course, this fidelity gain was at the cost of some additional CNOT gates and classical decoding steps.

V. DISCUSSION

A straightforward approach to achieving FTQC is to implement logical gates transversally, but no quantum stabilizer code can have a universal gate set of quantum computation that can be transversally implemented [39, 40]. As a consequence, universality is usually accomplished by the assistance of certain ancillas, prepared by magic state distillation [21]. A large literature is devoted to reducing the overhead of magic state distillation [41–46], since it may dominate the overall resources needed for FTQC [47].

The current work is motivated by a different approach to fault-tolerance: the use of a set of CSS codes to store and process quantum information. By teleporting logical qubits between code blocks that admit different sets of transversal gates, it is possible to perform a universal set of logical gates [48]. We have illustrated distillation with the simple example of the Steane $[[7,1,3]]$ code, but the procedure can readily be generalized to multi-qubit CSS codes. In particular, they can distill any CSS stabilizer state which is the simultaneous $+1$ eigenstate of all the stabilizer generators and a set of commuting logical operators, such that each logical operator includes the identity and either only X operators or only Z operators. This limitation means that it is not possible to distill completely general stabilizer states (or even general CSS stabilizer states) by the methods presented here, though generalizations of this scheme may make that possible. However, the protocols presented in this paper can distill all the logical ancillas needed for the teleportation-based FTQC scheme of [48].

Given the ability to prepare a set of suitable CSS stabilizer states, only transversal circuits and single-qubit Pauli measurements are needed for FTQC. In particular, magic states are not needed. (Other schemes without magic state distillation have also been proposed, such as [46, 49–52].) The results of this paper show that in principle this approach to FTQC is possible. However, the overhead for distillation dominates in this scheme, and we need to further analyze and quantify both the cost of distilling ancillas for various codes, and the performance of distillation in the presence of errors in the distillation circuit.

Our distillation protocol is a combination of the Steane syndrome extraction method and classical coding so that stabilizer states can be fault-tolerantly prepared. However, more work is needed to show that the overall scheme is both fault-tolerant and efficient enough to be useful at realistic error rates. In particular, noise in the distillation circuit will have two important effects. First, residual errors will be left in the ancilla by the noisy distillation circuit. Second, errors may degrade the performance so that ancilla errors from the encoding circuit may not be fully corrected. This first source of error will increase the effective error rate in the computation; but because the distillation circuit is transversal, these errors should be independent across the qubits of a single ancilla. However, the second type of error would be more serious: correlated errors across an ancilla can dramatically shorten the lifetime of the quantum codes used in the computation. Thus, careful modeling and numerical simulations are needed to assess and optimize the performance of distillation in the presence of noise.

Some methods may be used to greatly mitigate these potential pitfalls. In the distillation protocol by classical coding, error syndromes of the target ancillas are encoded by the coupling CNOTs and then recovered. If the transversal circuits for distillation are imperfect, the measured parity-check syndromes $\nu \mathsf{H}_Z^T$ are not reliable, which compromises the efficiency of distillation. However, this can be handled by learning more parities of the stabilizer generators as suggested by the method in [53–55]. In particular, we can choose another classical code $\mathcal{C}_3$ to encode the parity checks of H_Z by appending more redundant rows to the parity-check matrix. By calculating these additional parity checks, we can use any decoder of $\mathcal{C}_3$ to purify the decoding outputs of $\mathcal{C}_D$ and obtain more reliable error syndromes about the target ancillas. To further improve the reliability of distillation, we can also use a final postselection. By filtering out those noisy blocks with distinct syndromes, residual errors of higher weight can be further reduced at only a small cost in yield for the protocol.

We have begun to study these methods, and have found some preliminary results. A simulation of noisy distillation of the $[[23,1,7]]$ quantum Golay code by the $[3,1,3]$ repetition code, followed by the $[23,11,7]$ Golay code with postselection, suggests an overall yield of 10% by our protocol at a physical error rate of about $10^{-4} \sim 10^{-5}$. Technical details and performance analyses for some candidate classical codes and quantum codes will be addressed in a forthcoming paper [56].

In [12], an ancilla verification method is proposed for the quantum Golay code. This method can be regarded as a special case of our protocol, but our distillation protocol is potentially more efficient. In the verification scheme, pairs of blocks are compared to check errors. If any errors are detected in either block, everything is discarded and the process starts over again. By contrast, the results of each verification (parity checks) are kept track of in our protocol, and we can take advantage of their correlations by classical decoding. Since good classical codes with high rate and efficient decoders exist, we expect our distillation protocol to have higher throughput.

In the simulations in this paper, we used a minimum distance decoder of the classical code for distillation. However, the error rate of each syndrome bit of an ancilla may depend on the Clifford circuit that generates these faulty ancillas. We may be able to analyze this dependence and employ other techniques to improve the decoding performance. This is another future research direction.

Our distillation protocols are similar to magic state distillation, but there is an important distinction: because these ancillas are stabilizer states, they can be made using only Clifford gates, and (in principle) can be fault-tolerantly verified. This would suggest better performance here than in magic state distillation, where one cannot improve the quality of the encoded state by measuring it directly. At the very least, we should be able to do better in this respect: with magic state distillation, there is a probability of failure at each iteration step, where you have to discard everything; here, if we detect an error in the logical operators, we can correct them. Also, only certain codes with special properties can be used for magic state distillation; while a broad range of classical error-correcting codes can be applied in our scheme.

In this paper, we also have shown how to recover accurate error syndromes in Steane syndrome extraction using fewer ancillas, at the cost of higher classical decoding complexity and some additional CNOTs, while sacrificing a little channel fidelity. Since classical computing power is much cheaper, in general, than expensive quantum resources, it makes sense to exploit classical computing to save quantum resources. The layout of additional transversal CNOTs depends on the chosen classical code and their cost may or may not be comparable to the complexity saved by preparing fewer ancillas. However, the overall error-correcting power can be increased when the ancilla consumption rate is fixed. This protocol shares the same structure as ancilla distillation, and should give a net benefit at least in the regime of low error rates.

To do quantum error correction, we can also use the Shor syndrome extraction [3, 34]. (Knill syndrome extraction [58] is essentially equivalent to the Steane method up to qubit relocations.) For codes with low-weight stabilizer generators, Shor syndrome extraction may be preferred since it needs only low-weight ancillas—the cat states $(|0\rangle^{\otimes w} + |1\rangle^{\otimes w})/\sqrt{2}$—of size approximately equal to the weights of stabilizer generators. The cat states are also stabilizer states and thus could be prepared by our distillation protocols. We simply mentioned this since there are already methods of verifying the cat states.

Finally, our distillation protocols by classical error-correcting codes or quantum CSS codes could also be directly applied to the problem of multipartite entanglement purification for CSS stabilizer states. Conversely, a protocol for multipartite entanglement purification could potentially lead to a distillation protocol of CSS stabilizer states in FTQC if the multipartite constraint is removed. In particular, the purification protocols I/II in [23, 24] are a special case of our distillation protocol I by the $[2, 1, 2]$ classical error-detecting code with parity-check matrix

$$\mathsf{H}_D = \begin{pmatrix} 1 & 1 \end{pmatrix},$$

where an ancilla is discarded if the measurement outcomes are nonzero. However, we can use a general classical error-correcting code to do error recovery in the distillation process, and hence the efficiency can be better. Distillation protocol I could be adapted to a hashing protocol for multipartite entanglement purification as in [24–26]; however, we omit further discussion for the present, since the main topic of this paper is about fault-tolerant quantum computation. Also, the protocol in [28] is very similar to our distillation protocol II. Since they did not consider the problem in the fault-tolerant scenario, the eigenvalues of the stabilizers are measured directly in their protocol. By contrast, we use a transversal decoding circuit and bitwise qubit measurements to recover the eigenvalues of the stabilizers in our protocol.

ACKNOWLEDGMENTS

We thank Ben Reichardt for useful discussions. We also thank Scott Glancy for helpful comments. This work was supported in part by the IARPA QCS program; by HRL subcontract No. 1144-400707-DS; and by NSF Grant No. CCF-1421078. This work was also partially funded by the Singapore Ministry of Education (Tier-1 funding), through Yale-NUS Internal Grant IG14-LR001. T.A.B. also acknowledges funding as an IBM Einstein Fellow at the Institute for Advanced Study.

Appendix A: Justification of the Encoding for CSS Codes

The encoding procedure mentioned in Subsec. III B can be justified as follows. The unitary encoding operator can be implemented by applying a certain quantum circuit, consisting of CNOTs, Hadmard gates, phase gates and SWAP gates. (For example, Wilde gave an encoding algorithm [57] to find such a circuit.)

The check matrix of an $[[m, m-r-s]]$ CSS code can be written in the following form (see, e.g., [8]):

$$\mathsf{H} = [\mathsf{H}_X|\mathsf{H}_Z] = \left(\begin{array}{ccc|ccc} \mathsf{I}_r & \mathsf{A} & \mathsf{B} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{D} & \mathsf{I}_s & \mathsf{F} \end{array}\right),$$

where I_r and I_s are the $r \times r$ and $s \times s$ identity matrices, respectively, and A, B, D, and F are $r \times s$, $r \times (m-r-s)$, $s \times r$, and $s \times (m-r-s)$ binary matrices, respectively.

($r = s$ in our case.) Our goal is to apply a sequence of CNOT gates that transform H into

$$\left(\begin{array}{ccc|ccc} \mathsf{I}_r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathsf{I}_s & 0 \end{array}\right).$$

Then the reverse of this sequence of CNOTs is our encoding circuit.

This process is like applying Gaussian elimination on H. We first apply a series of CNOTs from the matrices I_r to clear the matrices A and B. These CNOTs have control qubits on qubit number 1 to r and target qubits on qubit number $r+1$ to m, respectively. Thus, only the matrix D of H_Z is altered by these CNOTs. We have

$$\mathsf{H}' = \left(\begin{array}{ccc|ccc} \mathsf{I}_r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{D}' & \mathsf{I}_s & \mathsf{F} \end{array}\right).$$

Since H' has to satisfy the commutation relations, D' must be 0. Thus we have

$$\mathsf{H}' = \left(\begin{array}{ccc|ccc} \mathsf{I}_r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathsf{I}_s & \mathsf{F} \end{array}\right).$$

Then we apply CNOTs to clear F. These CNOTs have control qubits on qubits number $(r+s+1)$ to m and target qubits on qubits number $(r+1)$ to $(r+s)$, respectively. H_X is not affected by these CNOTs and we have

$$\mathsf{H}'' = \left(\begin{array}{ccc|ccc} \mathsf{I}_r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathsf{I}_s & 0 \end{array}\right)$$

as required.

[1] D. Gottesman, Phys. Rev. A **57**, 127 (1998).

[2] D. Aharonov and M. Ben-Or, in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, STOC '97 (ACM, New York, NY, USA, 1997) pp. 176–188.

[3] D. P. DiVincenzo and P. W. Shor, Phys. Rev. Lett. **77**, 3260 (1996).

[4] F. Gaitan, *Quantum error correction and fault tolerant quantum computing* (CRC Press, Boca Raton, FL, 2008).

[5] D. A. Lidar and T. A. Brun, eds., *Quantum Error Correction* (Cambridge University Press, 2013).

[6] A. R. Calderbank and P. W. Shor, Phys. Rev. A **54**, 1098 (1996).

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

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

[9] P. Aliferis, D. Gottesman, and J. Preskill, Quant. Inf. Comp. **6**, 97 (2006).

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

[11] P. Aliferis and A. W. Cross, Phys. Rev. Lett. **98**, 220502 (2007).

[12] A. Paetznick and B. W. Reichardt, Quant. Inf. Comp. **12**, 1034 (2012).

[13] K. M. Svore, D. P. Divincenzo, and B. M. Terhal, Quant. Inf. Comp. **7**, 297 (2007).

[14] F. M. Spedalieri and V. P. Roychowdhury, Quant. Inf. Comp. **9**, 666 (2009).

[15] C.-Y. Lai, G. Paz, M. Suchara, and T. Brun, Quant. Inf. Comp. **14**, 807 (2014).

[16] A. G. Fowler, A. M. Stephens, and P. Groszkowski, Phys. Rev. A **80**, 052312 (2009).

[17] A. M. Steane, (2002) arXiv:quant-ph/0202036.

[18] H. Goto, Scientific Reports **6**, 19578 (2016).

[19] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Phys. Rev. A **54**, 3824 (1996).

[20] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Phys. Rev. Lett. **76**, 722 (1996).

[21] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).

[22] A. M. Steane, Phys. Rev. Lett. **78**, 2252 (1997).

[23] W. Dür, H. Aschauer, and H.-J. Briegel, Phys. Rev. Lett. **91**, 107903 (2003).

[24] H. Aschauer, W. Dür, and H.-J. Briegel, Phys. Rev. A **71**, 012319 (2005).

[25] K. Chen and H.-K. Lo, Quant. Inf. Comp. **7**, 689 (2007).

[26] E. Hostens, J. Dehaene, and B. De Moor, Phys. Rev. A **73**, 042316 (2006).

[27] C. Kruszynska, A. Miyake, H. J. Briegel, and W. Dür, Phys. Rev. A **74**, 052316 (2006).

[28] S. Glancy, E. Knill, and H. M. Vasconcelos, Phys. Rev. A **74**, 032319 (2006).

[29] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes* (North-Holland, Amsterdam, The Netherlands, 1977).

[30] M. Grassl, "Mathematics of quantum computation," (Chapman and Hall/CRC, 2002) Chap. Algorithmic aspects of quantum error-correcting codes.

[31] A. Córcoles, E. Magesan, S. J. Srinivasan, A. W. Cross, M. Steffen, J. M. Gambetta, and J. M. Chow, Nature communications **6** (2015).

[32] A. M. Steane, Proc. R. Soc. London A **452**, 2551 (1996).

[33] A. M. Steane, Nature **399**, 124 (1999).

[34] P. W. Shor, in *Proceedings of the 37th Annual Symposium on the Theory of Computer Science* (IEEE Press, Los Alamitos, 1996) pp. 56–65.

[35] C.-Y. Lai and C.-C. Lu, IEEE Trans. Inf. Theory **57**, 7163 (2011)

.

[36] B. Schumacher, Phys. Rev. A **54**, 2614 (1996).

[37] M. Reimpell and R. F. Werner, Phys. Rev. Lett. **94**, 080501 (2005).

[38] C.-Y. Lai and T. A. Brun, Phys. Rev. A **86**, 032319 (2012).

[39] B. Eastin and E. Knill, Phys. Rev. Lett. **102**, 110502 (2009).

[40] B. Zeng, A. Cross, and I. Chuang, IEEE Trans. Inf. Theory **57**, 6272 (2011).

[41] S. Bravyi and J. Haah, Phys. Rev. A **86**, 052329 (2012).

[42] B. Eastin, Phys. Rev. A **87**, 032321 (2013).

[43] C. Jones, Phys. Rev. A **87**, 022328 (2013).

[44] C. Jones, Phys. Rev. A **87**, 042305 (2013).

[45] C. Jones, Phys. Rev. A **87**, 052334 (2013).

[46] A. Paetznick and B. W. Reichardt, Phys. Rev. Lett. **111**, 090505 (2013).

[47] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Phys. Rev. A **86**, 032324 (2012).

[48] T. A. Brun, Y.-C. Zheng, K.-C. Hsu, J. Job, and C.-Y. Lai, (2015)arXiv:1504.03913.

[49] T. Jochym-O'Connor and R. Laflamme, Phys. Rev. Lett. **112**, 010505 (2014).

[50] J. T. Anderson, G. Duclos-Cianci, and D. Poulin, Phys. Rev. Lett. **113**, 080501 (2014).

[51] S. Bravyi, and A. Cross, (2015) arXiv:1509.03239.

[52] T. Jochym-O'Connor and S. D. Bartlett, Phys. Rev. A **93**, 022323 (2016).

[53] A. Ashikhmin, C.-Y. Lai, and T. A. Brun, in *Proc. IEEE Int. Symp. Inf. Theory* (2014) pp. 546–550.

[54] A. Ashikhmin, C.-Y. Lai, and T. A. Brun, in *Proc. IEEE Int. Symp. Inf. Theory* (2016) pp. 2274 - 2278.

[55] Y. Fujiwara, Phys. Rev. A **90**, 062304 (2014).

[56] Y.-C. Zheng, C.-Y. Lai, and T. A. Brun, Efficient Ancilla Distillation with Postselction by Classical Codes for Fault-tolerant Quantum Computation, (2016), in preparation.

[57] M. M. Wilde, *Quantum Coding with Entanglement*, Ph.D. thesis, University of Southern California (2008).

[58] E. Knill, (2003) arXiv:quant-ph/0312190.