Efficient preparation of large-block-code ancilla states for fault-tolerant quantum computation

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

Phys. Rev. A **97**, 032331 — Published 21 March 2018

DOI: 10.1103/PhysRevA.97.032331

Efficient preparation of large block code ancilla states for fault-tolerant quantum computation

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

[1]Centre for Quantum Technologies, National University of Singapore, Singapore, 117543
[2]Yale-NUS College, Singapore, 138527
[3]Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan
[4]Ming Hsieh Department of Electrical Engineering, Center for Quantum Information Science & Technology,
University of Southern California, Los Angeles, California 90089, USA

Fault-tolerant quantum computation (FTQC) schemes that use multi-qubit large block codes can potentially reduce the resource overhead to a great extent. A major obstacle is the requirement of a large number of clean ancilla states of different types without correlated errors inside each block. These ancilla states are usually logical stabilizer states of the data code blocks, which are generally difficult to prepare if the code size is large. Previously we have proposed an ancilla distillation protocol for Calderbank-Shor-Steane (CSS) codes by classical error-correcting codes. It was assumed that the quantum gates in the distillation circuit were perfect; however, in reality, noisy quantum gates may introduce correlated errors that are not treatable by the protocol. In this paper, we show that additional postselection by another classical error-detecting code can be applied to remove almost all correlated errors. Consequently, the revised protocol is fully fault-tolerant and capable of preparing a large set of stabilizer states sufficient for FTQC using large block codes. At the same time, the yield rate can be boosted from $O(t^{-2})$ to $O(1)$ in practice for an $[[n,k,d=2t+1]]$ CSS code. Ancilla preparation for the $[[23,1,7]]$ quantum Golay code is numerically studied in detail through Monte Carlo simulation. The results support the validity of the protocol when the gate failure rate is reasonably low. To the best of our knowledge, this approach is the first attempt to prepare general large block stabilizer states free of correlated errors for FTQC in a fault-tolerant and efficient manner.

I. INTRODUCTION

Quantum computers are difficult to build, since quantum states suffer from decoherence and quantum gates are imperfect [1]. However, the theory of quantum error-correcting codes (QECCs) [2–5] and fault-tolerant quantum computation (FTQC) [5–14] have shown that an arbitrarily large-scale quantum computation can be achieved if errors are not strongly correlated and their rates are small enough to fall below a threshold [7, 11–13, 15–17]. This is achieved by encoding the state of the computation in a quantum error-correcting code. Decoherence cannot corrupt an encoded state without affecting a large number of qubits.

Larger QECCs, with higher distances and rates, can potentially outperform smaller codes. Various FTQC schemes using multi-qubit large block codes have been proposed to achieve significantly higher code rates for comparable error protection ability [18–22]. For example, in Ref. [22], multiple logical qubits are encoded into high-rate large block codes and fault-tolerant (FT) logical gates are implemented by measuring logical states and teleporting logical qubits states back and forth between different code blocks to achieve universality, so that magic state distillation [23, 24] can be avoided. This can be done by preparation of different stabilizer ancilla states, transversal circuits, and bitwise qubit measurements.

High code rates and fault-tolerant non-Clifford gates without magic state distillation are two features that make the FTQC scheme based on large block codes promising. It potentially offers small resource overhead (see also [25–29] for other schemes with these features). However, ancilla states are required for logical teleportation and error correction. Straightforward preparation of these ancillas (e.g., through encoding circuits), is not fault-tolerant: a single circuit failure can result in a *high-weight* error in the output state with high probability. This is called a *correlated* error. As a result, the ancilla states need to be verified before they can be used.

The complexity of verifying prepared ancillas (in a straightforward way) grows quickly with code distance. For large codes, verification of encoded logical states such as $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ is accomplished by certain identically-prepared auxiliary ancillas. Errors from the target ancilla are copied into the auxiliary ancillas, which are then bitwise measured. If the measurement outcomes suggest that no error occurs, the ancillas are accepted. Otherwise, all the ancillas are discarded and the entire process restarts. Note that the auxiliary ancillas may also contain correlated errors after their preparation. These errors may propagate back to the target ancilla and the verification may fail even if the target ancilla is clean. Thus, the auxiliary ancillas must also be verified by more auxiliary ancillas. These recursive verifications will eventually dominate the overall resource overhead of quantum computation. In general, for an $[[n,k,d=2t+1]]$ CSS code, it requires $O(t^2)$ noisy ancilla states to generate a qualified one.

More efficient ways to prepare ancillas have been explored in the literature. Previously, verification of the ancilla state $|0\rangle_L$ has been studied for some quantum codes, such as the Steane code and the quantum Golay code [19, 30–33]. In Ref. [30], Steane proposed a method to filter out correct $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ for arbitrary Calderbank-Shor-Steane (CSS) codes [3, 34] using only single-round stabilizer measurements with postselection. However, it is highly likely that the overall rejection rate is too high to be of practical value when the code length is large. Moreover, that method can only remove one type (either X or Z) of correlated errors, which is good for the scheme in Ref. [18] but not sufficient for the one in Ref. [22]. In Ref. [33], the permutation symmetry of the quantum Go-

[*] zheng.yicong@quantumlah.org

lay code is exploited by permuting the qubits of different code blocks so that the correlation of errors between code blocks can be suppressed before verification. This simple operation can reduce the resource overhead of ancilla preparation by a factor of four. In Ref. [35], an efficient method was proposed to fault-tolerantly prepare $|0\rangle_L$ of the Steane code using only one extra qubit for verification. However, it is unclear how to generalize the above methods to arbitrary large block codes.

In Ref. [36], we proposed a general distillation protocol to prepare CSS stabilizer states, such as $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ states, for arbitrary CSS codes by classical error-correcting codes. The idea of the distillation protocol is as follows. Many ancilla states are identically prepared, each of which typically may contain highly correlated errors. They are fed into a distillation circuit, and certain check blocks are measured bitwise. The error syndromes of the remaining blocks can be extracted when the distillation circuit is based on the structure of a classical code. From the measurements, one can estimate the error syndromes of the target blocks by classical decoding. The protocol works correctly when the distillation circuit is perfect. However, if the CNOT gates in the distillation circuits are noisy, they can introduce correlated errors between different ancilla blocks so that classical decoding may not be able to recover these errors. Thus, the protocol is not completely fault-tolerant: it was assumed that the distillation circuit is perfect in Ref. [36], while in reality, noisy gates and measurements may lead to correlated errors after distillation.

In this paper we will modify the distillation protocol of Ref. [36] so that it works *fault-tolerantly* and *efficiently* when the distillation circuit is imperfect. The goal is for each block after fault-tolerant preparation to contain *no* correlated errors. Identifying the correct error syndromes for the ancillas is a key step in our distillation protocol. It is possible to introduce parity checks on the error syndromes by using another classical error-correcting code or an appropriate quantum stabilizer code [37–39], which greatly simplifies the process of syndrome verification. We apply a similar idea here to compute additional syndrome bits encoded by classical *error-detecting* codes. The estimated syndromes are checked to see whether they are compatible with the additional syndrome bits. If they agree, we accept the output; otherwise, they are discarded. After that, quantum error correction can be applied to the remaining blocks to remove both X and Z errors based on the estimated syndromes. Consequently, cleaner ancillas with no correlated errors can be obtained. The yield of the process depends on the rates of the chosen classical codes and the rejection rate of the output blocks. If an appropriate family of classical codes is chosen for postselection, the yield rate can be boosted from $O(t^{-2})$ to $O(1)$ in practice for an $[[n,k,d=2t+1]]$ CSS code.

Moreover, we prove that the output ancillas from this modified distillation protocol contain no correlated errors if sufficiently powerful classical codes are chosen. We prove a no-go theorem (Theorem 1) that a noisy distillation circuit will have correlated errors with high probability, which suggests that additional treatment is required when the distillation circuit is noisy. Our method of postselection by classical error-detecting codes is proved to be effective here (Theorem 4) when the underlying quantum code is *sparse* (see Def. 6 and Lemma 5) and the chosen classical code has error-correcting ability at least as good as the sparse quantum code (Theorems 2 and 3). This may even be true for general quantum codes (not necessarily sparse) and we demonstrate this by simulations of the distillation of the $[[23,1,7]]$ quantum Golay code in Sec. V.

In addition, we also propose protocols to distill various entangled ancilla states necessary for FTQC, such as the encoded EPR pairs between two code blocks. (The details are postponed to Appendix A.) In particular, we show how to do error correction and simultaneously measure a logical Y by Steane syndrome extraction.

The paper is organized as follows. We provide preliminary materials in Sec. II, including the basic ideas of quantum error-correcting codes, CSS codes and Steane syndrome extraction. In Sec. III, we review the protocol of Ref. [36] and formulate it for general multiple-qubit CSS codes. In Sec. IV, we propose a *fully* fault-tolerant protocol for these stabilizer ancilla states. In Sec. V, we numerically study the procedure for preparing $|0\rangle_L$ for the quantum Golay code by estimating the error weight distribution of the output ancillas for different combinations of small-sized classical codes. Discussion and suggestions for further improvements are presented in Sec. VI.

II. PRELIMINARIES

Both classical and quantum codes play essential roles in this paper. We start with a brief review of classical and quantum error-correcting codes and introduce our notation.

A. Classical Linear Codes

Classical error-correcting codes protect digital information by introducing redundancy. The encoded information strings—*codewords*— need to satisfy a set of *parity checks*, that is, linear constraints. Errors that cause violations of the parity checks can be detected and (hopefully) be corrected.

Let H be an $(n-k)\times n$ binary matrix. An $[n,k,d]$ classical binary code $\mathcal{C}$ associated with *parity-check matrix* H is a k-dimensional subspace of all binary $n-$tuples in $\mathbb{Z}_2^n$ such that

$$\mathsf{H}v^T = 0, \tag{1}$$

for all $v \in \mathcal{C}$, where v^T is the transpose of v and addition is modulo 2. Such vectors v are the codewords of $\mathcal{C}$. The parameter d is called the *minimum distance* of $\mathcal{C}$ such that any two codewords in $\mathcal{C}$ differ in at least d bits. This code can correct arbitrary $\lfloor \frac{d-1}{2} \rfloor$-bit errors.

For a received string $\tilde{v} \in \mathbb{Z}_2^n$, if $\mathsf{H}\tilde{v}^T \neq 0$, we know that some error occurred. Hence the rows of H are called *parity checks* of $\mathcal{C}$, and $\mathsf{H}\tilde{v}^T$ is called the *error syndrome* of $\tilde{v}$. A useful property of a linear code is that its parity-check matrix can be written in the systematic form:

$$\mathsf{H} = \left[\mathsf{I}_{n-k} | \mathsf{A}_{(n-k)\times k}\right]. \tag{2}$$

$\mathcal{C}$ can also be defined as the row space of a $k \times n$ generator matrix G, which satisfies

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

The dual code $\mathcal{C}^\perp$ of $\mathcal{C}$ is the row space of H. More properties of classical codes can be found in [40].

B. Stabilizer Codes and CSS Codes

The Hilbert space of a single qubit is the two-dimensional complex vector space $\mathbb{C}^2$ with an orthonormal basis $\{|0\rangle, |1\rangle\}$. Then the Hilbert space of $n-$qubit state is $\mathbb{C}^{2^n}$. Let $\mathcal{P}_n = \mathcal{P}_1^{\otimes n}$ denote the n-fold Pauli group, where

$$\mathcal{P}_1 = \{\pm I, \pm \mathrm{i} I, \pm X, \pm \mathrm{i} X, \pm Y, \pm \mathrm{i} Y, \pm Z, \pm \mathrm{i} Z\}, \tag{4}$$

and

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ Y = \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix},\ Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

For simplicity, we use the notation X_j to denote an X on qubit j, $I^{\otimes j-1} \otimes X \otimes I^{\otimes n-j}$, where n is the number of qubits. Y_j and Z_j are defined similarly. We also introduce the notation X^e for $e = e_1 \cdots e_n \in \mathbb{Z}_2^n$, to denote the operator $\otimes_{i=1}^n X^{e_i}$. An n-fold Pauli operator can be expressed as

$$\bigotimes_{i=1}^n X^{e_i} Z^{f_i} = X^e Z^f, \qquad e, f \in \mathbb{Z}_2^n. \tag{5}$$

up to an overall phase. Here (e, f) is called the *binary representation* of the Pauli operator $X^e Z^f$. We define the weight of E, wt(E), as the number of terms in the tensor product which are not equal to the identity.

Suppose $\mathcal{G}$ is an Abelian subgroup of $\mathcal{P}_n$ with a set of q independent and commuting generators $\{G_1, \ldots, G_q\}$, and $\mathcal{G}$ does not include $-I^{\otimes n}$. Every element in $\mathcal{P}_n$ has eigenvalues ± 1 or $\pm i$. An $[[n, k = n - q]]$ quantum stabilizer code $C(\mathcal{G})$ is defined as the 2^k-dimensional subspace of the n-qubit state space ($\mathbb{C}^{2^n}$) fixed by $\mathcal{G}$, which is the joint-$(+1)$ eigenspace of $G_1, \ldots, G_{n-k}$. Then for a codeword $|\psi\rangle \in C(\mathcal{G})$,

$$G|\psi\rangle = |\psi\rangle$$

for all $G \in \mathcal{G}$. When $k = 0$, $C(\mathcal{G})$ has only one eigenstate up to a global phase, and this state is called the *stabilizer state* of $\mathcal{G}$. Special stabilizer states play crucial roles in FTQC using large block codes [18, 19, 22], and the main object of this paper is to prepare these states fault-tolerantly.

If a Pauli error E corrupts $|\psi\rangle$, some eigenvalues of $G_1, \ldots, G_q$ will be flipped. Consequently, we gain error information by measuring the stabilizer generators $G_1, \ldots, G_q$, and the corresponding measurement outcomes (in bits), $g_1, \ldots, g_q$, are called the *error syndrome* of E. (The eigenvalue of a stabilizer generator is $+1$ or -1, and its corresponding syndrome bit is then 0 or 1, respectively. In the rest of the paper, the measurement outcomes are always represented in binary form.) A quantum decoder has to choose a good recovery operation based on the measured error syndromes.

CSS codes are an important class of stabilizer codes for FTQC. Their generators consist of tensor products of the identity and either X or Z operators (but not both) [3, 34]. More formally, consider two classical codes, $\mathcal{C}_Z$ and $\mathcal{C}_X$ with parameters $[n, k_Z, d_Z]$ and $[n, k_X, d_X]$, respectively, such that $\mathcal{C}_X^\perp \subset \mathcal{C}_Z$. The corresponding parity-check matrices are H_Z $(r_Z \times n)$ and H_X $(r_X \times n)$ with full rank $r_Z = n - k_Z$ and $r_X = n - k_X$. One can form an $[[n, k = k_X + k_Z - n, d]]$ CSS code $\mathcal{Q}$, where $d \geq \min\{d_Z, d_X\}$. For the special case that $\mathcal{C}_X = \mathcal{C}_Z$, we call such code a *symmetric* CSS code. In general a logical state of $\mathcal{Q}$ can be represented as:

$$|u\rangle_L = \sum_{x \in \mathcal{C}_X^\perp} |x + uD\rangle, \tag{6}$$

where $u \in \mathbb{Z}_2^k$ and D is a $k \times n$ binary matrix, whose rows are the coset leaders of $\mathcal{C}_Z/\mathcal{C}_X^\perp$. The Z and X stabilizer generators of $\mathcal{Q}$ are

$$G_{Z|i} = \bigotimes_{j=1}^n Z^{[\mathsf{H}_Z]_{i,j}}, \qquad i = 1, \ldots, r_Z, \tag{7}$$

and

$$G_{X|i} = \bigotimes_{j=1}^n X^{[\mathsf{H}_X]_{i,j}}, \qquad i = 1, \ldots, r_X. \tag{8}$$

Here $[\mathsf{M}]_{i,j}$ represents the (i, j) entry of a matrix M. The *check matrix* for $\mathcal{Q}$ is defined as

$$\mathsf{H} = \begin{bmatrix} \mathsf{H}_Z & 0 \\ 0 & \mathsf{H}_X \end{bmatrix}. \tag{9}$$

The error syndrome of a Pauli error is a binary string of outcomes of measuring the stabilizers $G_1, \ldots, G_{n-k}$. According to Eq. (5), error correction can be done by treating X and Z errors separately. For CSS codes, the eigenvalues of $G_{Z|1}, \ldots, G_{Z|r_Z}$ $(G_{Z|1}, \ldots, G_{Z|r_X})$ correspond to the error syndrome of X (Z) errors. Then its X and Z error syndromes $g_Z \in \mathbb{Z}_2^{r_Z}$, $g_X \in \mathbb{Z}_2^{r_X}$ are given by

$$g_Z^T = \mathsf{H}_Z e^T, \quad g_X^T = \mathsf{H}_X f^T. \tag{10}$$

The Pauli operators that are not in $\mathcal{G}$ but commute with all the stabilizer generators are called the *logical Pauli operators*. For the CSS codes defined in Eq. (6), it is always possible to find a subset of logical operators involving only tensor product of identity and either X or Z operators. Explicitly, the logical X operators for the jth logical qubit $\bar{X}_j$ can be written as

$$\bar{X}_j = \bigotimes_{i=0}^n X_i^{[D]_{j,i}}. \tag{11}$$

The logical Z operators for the jth logical qubit can be constructed as [3, 18]:

$$\bar{Z}_j = \bigotimes Z_i^{[(DD^T)^{-1}D]_{j,i}}. \tag{12}$$

C. Steane Syndrome Extraction

For a CSS code $\mathcal{Q}$, Steane suggested a method to extract syndromes as shown in Fig. 1 [41]. This method needs two clean ancilla states, which are called the X and Z ancillas, since they will be bitwise measured in the X and Z bases, respectively. If the X and Z ancillas are in the logical states $|0\rangle_L$ and $|+\rangle_L = \frac{1}{\sqrt{2}}(|0\rangle_L + |1\rangle_L)$ of $\mathcal{Q}$, the circuit extracts the error syndromes without disturbing the encoded quantum information. If other ancilla states are used, it is possible also to measure a logical Pauli operator on the data block. Since a logical gate can be implemented by measuring a sequence of logical operators [22], this leads to a fault-tolerant implementation of logical gate operation and syndrome extraction, simultaneously. As we can see in Appnedix. A 1, these ancillas are stabilizer states of $\mathcal{Q}$. Namely, they are stabilized by both the stabilizer generators of $\mathcal{G}$ and some logical Pauli operators. For a logical ancilla of a single code block, its corresponding set of stabilizers is:

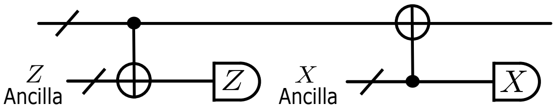

FIG. 1. Quantum circuit for Steane syndrome extraction.

$$S = \{G_{X|1}, \ldots, G_{X|r_X}, G_{Z|1}, \ldots, G_{Z|r_Z}, L_1, \cdots, L_k\}. \tag{13}$$

where the L_i are logical Pauli operators that commute with each other. Note that we may also need ancilla states that are entangled among $m > 1$ different blocks. The stabilizer generators and logical Pauli operators of such a logical ancilla can be represented by:

$$\begin{aligned} S = \{&G^{(1)}_{X|1}, \ldots, G^{(1)}_{X|r_X}, G^{(1)}_{Z|1}, \ldots, G^{(1)}_{Z|r_Z}, \ldots \\ &G^{(m)}_{X|1}, \ldots, G^{(m)}_{X|r_X}, G^{(m)}_{Z|1}, \ldots, G^{(m)}_{Z|r_Z}, L_1, \ldots, L_{mk}\} \end{aligned} \tag{14}$$

Two simple examples are the logical $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ states, where all logical qubits are in the states $|0\rangle_L$ and $|+\rangle_L$, respectively. These two states are stabilized by

$$S = \{G_{X|1}, \ldots, G_{X|r_X}, G_{Z|1}, \ldots, G_{Z|r_Z}, \bar{Z}_1, \ldots, \bar{Z}_k\}, \tag{15}$$

and

$$S = \{G_{X|1}, \ldots, G_{X|r_X}, G_{Z|1}, \ldots, G_{Z|r_Z}, \bar{X}_1, \ldots, \bar{X}_k\}, \tag{16}$$

respectively. In the rest of the paper, these stabilizer generators and logical Pauli operators will be called *stabilizers* for short. The measurement outcomes of their eigenvalues in binary form are called *generalized syndromes* for simplicity.

D. Failures and Errors

For clarification, we need to distinguish between *failures* and *errors*. Failures are physical processes which cause imperfection on certain qubits at particular time in the circuit. The resulting effect of failures on the quantum state are called *errors*. In general, a single-qubit failure in the circuit can result in multiple-qubit errors.

A basic assumption here is that failures in different places and times of a quantum circuit are *independent*. This assumption can be relaxed, but for simplicity, we retain it for this paper. In this paper, we consider only Pauli failures. Assume that at each time step, every physical qubit independently undergoes an X, Y or Z error with probability $\epsilon/3$. Such failure is called a *memory failure*. For a CNOT gate, failure is modeled as a perfect gate followed by one of the 15 possible failures from IX, IY , IZ, XI, XX, XY , XZ, YI, YX, YY , YZ, ZI, ZX, ZY, and ZZ with equal probability $p_g/15$. We call this a *gate failure*. Also, a measurement of a single physical qubit suffers a classical bit-flip error with probability p_m (an X or Z error preceding a measurement in the Z or X basis, respectively). We call this a *measurement failure*. In the rest of the paper, we assume that the gate failure rate is equal to the measurement failure rate—i.e., $p_g = q_m = p$. Also, we consider the scenario where the memory failure rate is much smaller than the gate failure rate, that is, $\epsilon \ll p$, so that it can be safely neglected. This is usually the case for a physical system with fast gates and long coherence time, like ion trap [42] or superconducting qubits [43].

Correlated errors are usually very harmful for FTQC. The following definition is based on Ref. [30]:

Definition 1. *Consider a quantum code $\mathcal{Q}$ which can correct any Pauli error on t qubits. We say that an error E is* uncorrelated *if the probability of E after the preparation of a codeword of $\mathcal{Q}$ scales like*

$$\Pr(E) \sim O(p^s) : \begin{cases} \textit{for some } s \geq wt(E), \ \textit{if } wt(E) \leq t; \\ \textit{for some } s \geq t, \ \textit{if } wt(E) > t. \end{cases} \tag{17}$$

Otherwise, E is said to be *correlated*. The parameter s is called the *order* of error E. Note that the coefficients behind the big O notation should not be unreasonably large. Note also that, for uncorrelated errors, the probability distribution of their weights is somewhat like a binomial distribution for error order less than t.

To realize the FTQC protocol in [18, 19, 22], the ancillas should contain no correlated errors, so that correlated errors will not propagate from the ancillas to the data block, introducing uncorrectable errors. Even if errors on an ancilla are correctable, if their order is not strictly greater than or equal to their weights, then they can propagate to the data blocks and cause quantum error correction to fail in the coming Steane syndrome extraction with much higher probability then if they were uncorrelated. We say an ancilla is *qualified* if it is free of correlated errors. If both Z and X ancillas are qualified, then single-shot Steane syndrome extraction is sufficient to implement a logical gate operation or a logical state measurement [22].

III. ANCILLA PREPARATION USING CLASSICAL CODES WITH NOISELESS QUANTUM CIRCUITS

Preparing these states with high quality and high throughput is crucial. Here is a naive preparation procedure of these ancillas, through verification using an additional identically prepared ancilla state. Errors are copied from the initial ancilla to an auxiliary ancilla, which are then bitwise measured. If errors are detected, one discards both ancillas. Note that the auxiliary ancillas may also contain correlated errors after their preparation, and need further checks by more auxiliary ancillas. This procedure ultimately would will involve tremendous numbers of ancilla states.

For example, the verification circuit for a qualified ancilla $|0\rangle_L$ of the $[[23,1,7]]$ Golay code is shown in Fig. 2 [33]. In general, to verify a t−error-correcting quantum code, one needs a distillation process using at least t^2+t identically prepared blocks [44] , which can be very large for the codes of interest. Denote the overall rejection rate of naive verification by $R_{\text{naive}}(p)$, which is $O(p)$, since any single failure will lead to a rejection. The average yield rate of qualified ancillas is:

$$\text{Yield} = \frac{1 - R_{\text{naive}}(p)}{t^2+t}. \tag{18}$$

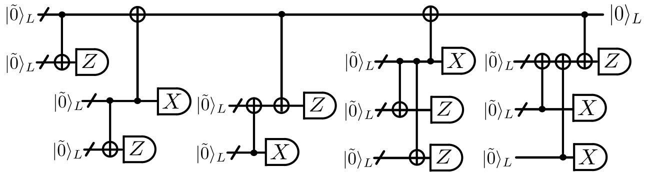

FIG. 2. A naive circuit that produces a $|0\rangle_L$ state of the $[[23,1,7]]$ Golay code free of correlated errors. The CNOTs are transversal and measurements are bitwise. All $|\tilde{0}\rangle_L$ are identically prepared. If any measurement suggests the existence of errors, all ancilla blocks involved are discarded and the procedure restarts.

In other contexts (e.g., entanglement purification [45, 46]) we may get around this formidable verification procedure by *distillation*, which takes many of imperfect states and carries out a protocol to produce a smaller number of better states. In Ref. [36], we have shown that classical error-correcting codes can be applied to distill stabilizer states, such as $|0\rangle_L$ and $|+\rangle_L$, if the distillation circuit is perfect. When errors are present in blocks, we can correct them in most cases, rather than discarding the ancillas and starting over. Potentially, this scheme can greatly reduce the resource overhead.

In this section, we review distillation for CSS stabilizer states $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ using classical error-correcting codes with a *noiseless* quantum circuit [36]. We will formalize the protocol for general $[[n,k,d]]$ CSS codes. Generalization of the protocol to handle all necessary ancilla states for FTQC can be found in Appendix. A 2. The problem of noisy distillation circuits will be discussed in detail in the next section.

We consider the ancilla states $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ of an $[[n,k,d]]$ CSS code $\mathcal{Q}$ constructed from an $[n,k_X,d_X]$ code $\mathcal{C}_X$ and an $[n,k_Z,d_Z]$ code $\mathcal{C}_Z$ such that $\mathcal{C}_X^\perp \subset \mathcal{C}_Z$. The distillation protocol in [36] consists of two rounds of distillation, where two classical code $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ are used to correct X and Z errors, respectively. Consider an $[n_{c_1},k_{c_1},d_{c_1}]$ classical code $\mathcal{C}_{c_1}$ and an $[n_{c_2},k_{c_2},d_{c_2}]$ classical codes $\mathcal{C}_{c_2}$ that can correct $t_{c_1} = \lfloor\frac{d_{c_1}-1}{2}\rfloor$ and $t_{c_2} = \lfloor\frac{d_{c_1}-1}{2}\rfloor$ errors, respectively. Let $\mathsf{H}_{c_1} = [\mathsf{I}_{r_{c_1}} | \mathsf{A}_{c_1}]$ be the parity-check matrix of $\mathcal{C}_{c_1}$ in systematic form, where $r_{c_1} = n_{c_1} - k_{c_1}$, $\mathsf{I}_{r_{c_1}}$ is the $r_{c_1} \times r_{c_1}$ identity, and A_{c_1} is an $r_{c_1} \times k_{c_1}$ binary matrix.

Suppose we have $n_{c_1} n_{c_2}$ independently prepared ancillas in the $|0\rangle_L^{\otimes k}$ state (for $|+\rangle_L^{\otimes k}$ state, the situation is similar). This can be done by an encoding circuit using CNOTs and physical state preparations in $|0\rangle$ and $|+\rangle$. Note that these process are noisy, so after encoding, for each block, there exists a residual Pauli error E of the form $X^e Z^f$. We then divide them into n_{c_2} groups of n_{c_1} blocks. For each group of n_{c_1} noisy ancilla blocks, let $g^{(j)}_{Z|i}$ be the eigenvalue of the stabilizer generators $G_{Z|i}$ of the jth block, and $\ell^{(j)}_{\bar{Z}_i}$ be the eigenvalue of $\bar{Z}_i$ for the jth block. Then the syndromes of stabilizers for the n_{c_1} ancilla blocks of $\mathcal{Q}$ are:

$$\begin{array}{cccc}
g^{(1)}_{Z|1} & g^{(1)}_{Z|2} & \cdots & g^{(1)}_{Z|r_Z-1} & g^{(1)}_{Z|r_Z} \\
g^{(2)}_{Z|1} & g^{(2)}_{Z|2} & \cdots & g^{(2)}_{Z|r_Z-1} & g^{(2)}_{Z|r_Z} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
g^{(n_{c_1}-1)}_{Z|1} & g^{(n_{c_1}-1)}_{Z|2} & \cdots & g^{(n_{c_1}-1)}_{Z|r_Z-1} & g^{(n_{c_1}-1)}_{Z|r_Z} \\
g^{(n_{c_1})}_{Z|1} & g^{(n_{c_1})}_{Z|2} & \cdots & g^{(n_{c_1})}_{Z|r_Z-1} & g^{(n_{c_1})}_{Z|r_Z},
\end{array} \tag{19}$$

and

$$\begin{array}{ccccc}
\ell^{(1)}_{\bar{Z}_1} & \ell^{(1)}_{\bar{Z}_2} & \cdots & \ell^{(1)}_{\bar{Z}_{k-1}} & \ell^{(1)}_{\bar{Z}_k} \\
\ell^{(2)}_{\bar{Z}_1} & \ell^{(2)}_{\bar{Z}_2} & \cdots & \ell^{(2)}_{\bar{Z}_{k-1}} & \ell^{(2)}_{\bar{Z}_k} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\ell^{(n_{c_1}-1)}_{\bar{Z}_1} & \ell^{(n_{c_1}-1)}_{\bar{Z}_2} & \cdots & \ell^{(n_{c_1}-1)}_{\bar{Z}_{k-1}} & \ell^{(n_{c_1}-1)}_{\bar{Z}_k} \\
\ell^{(n_{c_1})}_{\bar{Z}_1} & \ell^{(n_{c_1})}_{\bar{Z}_2} & \cdots & \ell^{(n_{c_1})}_{\bar{Z}_{k-1}} & \ell^{(n_{c_1})}_{\bar{Z}_k}.
\end{array} \tag{20}$$

Let S_1 be the set of $g^{(j)}_{Z|i}$ and $\ell^{(j)}_{\bar{Z}_i}$, which form a generalized syndrome array:

$$[\mathsf{s}_1]_{j,i} = \begin{cases} g^{(j)}_{Z|i}, & \text{for } 1 \le i \le r_Z; \\ \ell^{(j)}_{\bar{Z}_{i-(n-n_Z)}}, & \text{for } r_Z < i \le r_Z + k. \end{cases} \tag{21}$$

Here, the subscript 1 means that the eigenvalues of these stabilizers are processed at the first round of distillation. Similarly, we define S_2 as the set of X stabilizer generators whose syndrome array s_2 would be estimated at the second round of distillation. Obviously, $S_1 \cup S_2$ is the complete set of stabilizer generators.

If the preparation circuits are perfect, the entries of s_1 would all be 0s. However, if circuits are imperfect, some generalized syndromes bits will be flipped to 1s. Let $[\mathsf{s}_1]_{:,i}$ denote the ith column of s_1. If one can correctly identify a row j of $[\mathsf{s}_1]$, $[\mathsf{s}_1]_{j,:}$, i.e., the generalized syndromes for the jth block, then in principal quantum error correction can be carried out to correct the errors on the jth block. Observe that the entries of $[\mathsf{s}_1]_{:,i}$ are independent binary random variables, since the ancilla blocks are prepared independently. In

Ref. [36], we introduce the idea of distillation by classical codes, so that a subset of $[\mathsf{s}_1]_{:,i}$ can be determined.

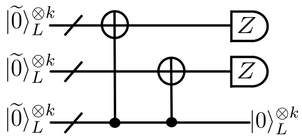
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
Z
Z
$|0\rangle_L^{\otimes k}$

FIG. 3. The circuit for ancilla distillation by the classical $[3,1,3]$ repetition code to remove X errors. The first two $|\widetilde{0}\rangle_L^{\otimes k}$ serve as parity-check ancillas.

Choose the first r_{c_1} of the ancillas of each group to hold the classical parity checks, and do transversal CNOTs from the remaining k_{c_1} ancillas onto each of the parity-check ancillas according to the pattern of 1s in the rows of A_{c_1}. If $[\mathsf{A}_{c_1}]_{i,j} = 1$, we apply a transversal CNOT from the $(r_{c_1}+j)$th ancilla to the ith ancilla. Note that X errors will not propagate from the ith block to $(r_{c_1}+j)$th block through the CNOTs applied. Then measure all the qubits on each of the first r_{c_1} ancilla blocks in the Z basis, which destroys the states of those blocks. As an example, Fig. 3 demonstrates the distillation circuit using the $[3,1,3]$ repetition code. From the measurement outcomes $\nu^{(1)},\cdots,\nu^{(r_{c_1})} \in \mathbb{Z}_2^n$, we obtain r_{c_1} strings $\sigma^{(i)} \in \mathbb{Z}_2^{r_Z+k}$, $i = 1,\ldots,r_{c_1}$, which can be shown to be

$$\mathsf{H}_{c_1}\mathsf{s}_1 = \begin{bmatrix} \nu^{(1)}\left(\mathsf{H}'_Z\right)^T \\ \vdots \\ \nu^{(r_{c_1})}\left(\mathsf{H}'_Z\right)^T \end{bmatrix} \equiv \begin{bmatrix} \sigma^{(1)} \\ \vdots \\ \sigma^{(r_{c_1})} \end{bmatrix} \equiv \sigma, \qquad (22)$$

where

$$\mathsf{H}'_Z = \begin{bmatrix} \mathsf{H}_Z \\ \mathsf{L}_Z \end{bmatrix} \equiv \begin{bmatrix} \mathsf{H}_Z \\ \hline [\bar{z}_1^T \cdots \bar{z}_k^T]^T \end{bmatrix}, \qquad (23)$$

and $\bar{z}_j$ is the binary representation of $\bar{Z}_j$. In other words, the ith column of σ matrix has the generalized syndrome bits of $[\mathsf{s}_1]_{:,i}$. Then we can use a decoder for $\mathcal{C}_{c_1}$ to estimate the values of $[\mathsf{s}_1]_{:,i}$ for all i and thus obtain the estimated generalized syndrome array $[\widetilde{\mathsf{s}}_1]$. After that, for code block $r_{c_1}+1 \le j \le n_{c_1}$, one can use $\{\tilde{g}^{(j)}_{Z|i}\}$ and a classical decoder for $\mathcal{C}_Z$ to estimate the X errors. Suppose the decoder of $\mathcal{C}_Z$ always generates an estimate $\tilde{e}$ of an error vector e such that $\mathsf{H}_Z(e+\tilde{e}) = 0$, and assume the decoded X error on block j is $E_X^{(j)}$. Then for each logical operator $\bar{Z}_i^{(j)}$, $i = 1,\ldots,n$, $j = r_{c_1}+1,\ldots,n_{c_1}$, follow the rules:

1. If $[E_X^{(j)}, \bar{Z}_i^{(j)}] = 0$ and $\ell^{(j)}_{\bar{Z}_i} = 0$, then do nothing;
2. If $\{E_X^{(j)}, \bar{Z}_i^{(j)}\} = 0$ and $\ell^{(j)}_{\bar{Z}_i} = 0$, then replace $E_X^{(j)}$ by $E_X^{(j)}\bar{X}_i^{(j)}$;
3. If $[E_X^{(j)}, \bar{Z}_i^{(j)}] = 0$ and $\ell^{(j)}_{\bar{Z}_i} = 1$, then replace $E_X^{(j)}$ by $E_X^{(j)}\bar{X}_i^{(j)}$;
4. If $\{E_X^{(j)}, \bar{Z}_i^{(j)}\} = 0$ and $\ell^{(j)}_{\bar{Z}_i} = 1$, then do nothing.

The updated $E_X^{(j)}$ are the decoded X errors, and one can correct them on the output block or just keep track of them. Alternatively, for a symmetric CSS code, we may decode $\mathcal{C}_Z^\perp$ directly from $[\widetilde{\mathsf{s}}_1]_{j,:}$ to get $E_X^{(j)}$ directly. However it may be very difficult to find an efficient decoder for $\mathcal{C}_Z^\perp$ in practice.

Suppose the gate failure rate in the encoding circuit of $|0\rangle_L^{\otimes k}$ is p. A gate failure may propagate through in the whole circuit so that, after encoding, the ancilla block contains some X and Z errors of weight up to the code length with probability $O(p)$. After the distillation process removes X errors, the remaining k_{c_1} ancillas from each group will have lower rates of X errors than they started with. The probability that X errors of arbitrary weight remain on each block drops from $O(p)$ to $O(cp^{t_{c_1}+1})$, where c is a constant that depends on the details of $\mathcal{C}_{c_1}$ and its decoder. This is because the estimated generalized syndromes are wrong with probability $O(p^{t_c+1})$. But correlated Z errors on the target qubits can propagate via the CNOTs back onto the these k_{c_1} blocks. The rate of Z errors on the remaining k_{c_1} ancillas will increase to $O((M+1)p)$, where M is the maximum number of parity checks each control qubit is included in (or equivalently, the maximum number of 1s in each column of the matrix A_{c_1}). If p is not too large, the rate of Z errors has grown roughly by a constant factor M, while the rate of X errors has been substantially reduced. From the original large number of ancillas, we retain a fraction k_{c_1}/n_{c_1}, which have much lower X error rates and somewhat higher Z error rates.

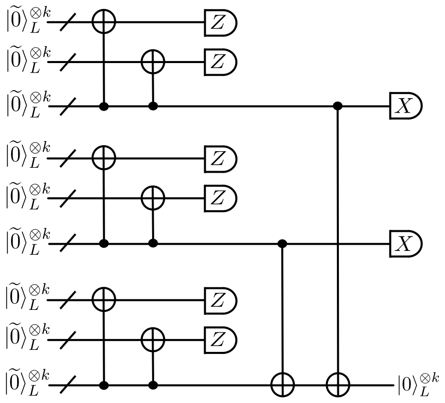
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
$|\widetilde{0}\rangle_L^{\otimes k}$
Z
Z
X
Z
Z
X
Z
Z
$|0\rangle_L^{\otimes k}$

FIG. 4. The circuit for ancilla distillation by the classical $[3,1,3]$ repetition code to produce ancillas free of (low-weight) X and Z errors.

Each group will output k_{c_1} blocks after the first round of distillation. To remove the remaining Z errors, we divide the $k_{c_1}n_{c_2}$ output blocks into k_{c_1} groups of n_{c_2} blocks. Note that here it is important that ancillas that were grouped together in the first round *not* be grouped together in the second round, because their Z errors are now correlated between different blocks in the same group. This randomization process will ensure that the flipping of the generalized syndromes for different blocks in each group are independent so that classical error correction can work properly. Similarly, we can use the classical code $\mathcal{C}_{c_2}$ to remove Z errors. The process is similar to the first found, each qubit of the first $r_{c_2} = n_{c_2} - k_{c_2}$ ancillas is measured in the X basis to estimate $\{\tilde{g}^{(j)}_{X|i}\}$ by classical decoding of $\mathcal{C}_{c_2}$. (For $|0\rangle_L$, the eigenvalues of $\bar{X}_i$ do not need to be evaluated.) Since the error rate for X is now higher, one may choose $\mathcal{C}_{c_2}$ with a stronger error correction ability than $\mathcal{C}_{c_1}$. An example of the whole distillation procedure using the $[3,1,3]$ repetition code is shown in Fig. 4. At the end, the overall yield rate is $\frac{k_{c_1}k_{c_2}}{n_{c_1}n_{c_2}}$.

IV. FAULT-TOLERANT ANCILLA STATE PREPARATION USING CLASSICAL CODES

In the FTQC, CNOTs and measurements in a distillation circuit can be noisy, which could destroy the whole protocol, producing unqualified ancillas as outputs. In this section, we will first study the negative effects of noisy circuits on the output ancillas, and show that the previous distillation protocol fails in this situation. We will then modify the distillation protocol to be a fully fault-tolerant, and show that the output ancillas are qualified.

In general, a fault-tolerant ancilla preparation procedure has two stages: ancilla preparation by noisy circuits and ancilla distillation by noisy circuits. Before we go on, we introduce the notation that will be used later. Let $\mathscr{G}$ be the set of failures in the whole procedure and $|\mathscr{G}|$ represents the number of failures. Note that we only consider failures that are Pauli operators. Hence any single failure $\mathrm{g} \in \mathscr{G}$ can be decomposed as a product of an X part and a Z part, up to a phase: $\mathrm{g} = \mathrm{g}_X \mathrm{g}_Z$. We define the collection of g_X's (g_Z's) from $\mathscr{G}$ as $\mathscr{G}_X$ ($\mathscr{G}_Z$). Obviously, $|\mathscr{G}_X|, |\mathscr{G}_Z| \leq |\mathscr{G}|$.

We can define a subset $\mathscr{G}(\tau) \subseteq \mathscr{G}$, which contains failures that occur at the τth step of the circuit. Let $\mathrm{Q}_{\mathscr{G}(\tau)}$ be the *support* of $\mathscr{G}(\tau)$. For example, for a single CNOT gate at the τth step, if $\mathrm{g} \in \mathscr{G}(\tau)$ acts nontrivially on the control (target) qubit, then $\mathrm{Q}_{\mathscr{G}(\tau)}$ contains the control (target) qubit; if $\mathrm{g} \in \mathscr{G}(\tau)$ affects both the control and target qubits, then $\mathrm{Q}_{\mathscr{G}(\tau)}$ contains both qubits. Let $\mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}$ denote the support of failures on block j at the τth step. Then $\mathrm{Q}_{\mathscr{G}(\tau)} = \bigcup_j \mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}$.

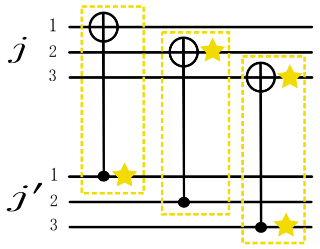

FIG. 5. (Color online) A single step in the preparation circuit involving two blocks j and j'. The yellow stars represent the supporting qubits of failures that occur on the CNOT gates.

We also define $\mathcal{I}\left(\mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}\right) \subseteq \{1,\ldots,n\}$, where n is the code length, as the set of qubit indices of $\mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}$ in block j. For example, consider a single step in the preparation procedure as shown in Fig. 5. The corresponding $\mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}$ and $\mathrm{Q}^{(j')}_{\mathscr{G}(\tau)}$ are qubits 2, 3 of block j and qubit 1, 3 of block j'.

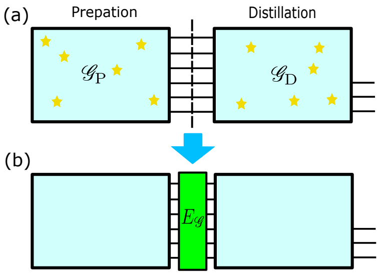

FIG. 6. (Color online) (a) Failure set $\mathscr{G}$ (whose supports are represented by yellow stars) can be decomposed into failures occurring in the preparation stage $\mathscr{G}_{\mathrm{P}}$ and in the distillation stage $\mathscr{G}_{\mathrm{D}}$. (b) The effect of $\mathscr{G}$ is equivalent to perfect ancilla preparation and perfect distillation circuits with an effective failure process in between.

Our strategy is to transform this fault-tolerant ancilla preparation procedure into one with three stages: 1) perfect ancilla preparation; 2) an effective failure process; and 3) perfect distillation circuits. Thus $\mathscr{G}$ can be decomposed as $\mathscr{G} = \mathscr{G}_{\mathrm{P}} \bigcup \mathscr{G}_{\mathrm{D}}$, where $\mathscr{G}_{\mathrm{P}}$ and $\mathscr{G}_{\mathrm{D}}$ are failure sets in the preparation and distillation stages, respectively. (See, for example, Fig. 6.) It is always possible for a noisy preparation circuit to be modeled as a perfect one followed by some failures, and a distillation circuit can be modeled as a perfect one with preceding failures. Let

$$E_{\mathscr{G}} = \bigcup_j E^{(j)}_{\mathscr{G}}$$

be the effective failures of $\mathscr{G}$ at the second stage, where $E^{(j)}_{\mathscr{G}}$ is set of the effective failures on block j. One can also define $\mathrm{Q}^{(j)}_{E_{\mathscr{G}}}$, $\mathcal{I}\left(\mathrm{Q}^{(j)}_{E_{\mathscr{G}}}\right)$ in a similarly way. Define $S^{(j)}_{\mu}$ as the stabilizers of block whose eigenvalues need to be estimated in the μth round of distillation, for $\mu = 1, 2$. Let $S^{(j)}_{\mu}|_{E_{\mathscr{G}}} \subseteq S^{(j)}_{\mu}$ be the subset of stabilizers anticommuting with $E_{\mathscr{G}}$ on block j (the set $E_{\mathscr{G}}$ is a Pauli operator) and $\mathcal{I}\left(S^{(j)}_{\mu}|_{E_{\mathscr{G}}}\right)$ be the set of indices of stabilizers in $S^{(j)}_{\mu}|_{E_{\mathscr{G}}}$.

A. A no-go theorem for the previous protocol with noisy circuits

Observe that in a distillation circuit, the qubits that are measured in the Z basis will never be the control qubits of CNOT gates. If a noisy measurement is modeled as a perfect measurement with a preceding random Pauli X failure, this X failure will not propagate through the CNOTs. Hence, this type of X failure on the measurement can be directly absorbed into the effective failure preceding a perfect distillation. Similarly for the Z failures before the measurements in the X basis. Thus, measurement failures only have a very limited effect. However, the effect of noisy CNOTs in the distillation circuit is more complicated and will be discussed as follows.

1. Perfect preparation circuit with noisy distillation circuits

First, we consider the case that the input ancilla blocks are free of errors, while the distillation circuits are noisy. This seems absurd at a first glance, since then no distillation is needed. Nevertheless, it will give insight about the distribution of errors caused by the noisy CNOTs. For such a restricted set of failures $\mathscr{G}$, the *effective* support of $\mathscr{G}_X$ on block j is defined as

$$\mathrm{QE}^{(j)}_{\mathscr{G}_X} = \bigsqcup_{\tau} \mathrm{Q}^{(j)}_{\mathscr{G}_X(\tau)}, \tag{24}$$

where the operator $\bigsqcup$ takes the union of all $\mathrm{Q}^{(j)}_{\mathscr{G}_X(\tau)}$ except those qubits affected by an even number of different elements in $\mathscr{G}_X(\tau)$. For example, consider two CNOT gates with the same control qubit. Suppose that each of the two CNOTs suffers an X failure on the control qubit, and $\mathscr{G}_X$ contains only these two failures. Then $\mathrm{QE}_{\mathscr{G}_X} = \varnothing$ since the two X failures will cancel each other out. $\mathrm{QE}^{(j)}_{\mathscr{G}_X}$ is similarly defined.

Recall that the whole array of generalized syndromes before the μth round of distillation can be listed as follows:

$$\begin{matrix} [\mathsf{s}_\mu]_{1,1} & [\mathsf{s}_\mu]_{1,2} & \cdots & [\mathsf{s}_\mu]_{1,|S_\mu|-1} & [\mathsf{s}_\mu]_{1,|S_\mu|} \\ [\mathsf{s}_\mu]_{2,1} & [\mathsf{s}_\mu]_{2,2} & \cdots & [\mathsf{s}_\mu]_{2,|S_\mu|-1} & [\mathsf{s}_\mu]_{2,|S_\mu|} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ [\mathsf{s}_\mu]_{n_{c_\mu}-1,1} & [\mathsf{s}_\mu]_{n_{c_\mu}-1,2} & \cdots & [\mathsf{s}_\mu]_{n_{c_\mu}-1,|S_\mu|-1} & [\mathsf{s}_\mu]_{n_{c_\mu}-1,|S_\mu|} \\ [\mathsf{s}_\mu]_{n_{c_\mu},1} & [\mathsf{s}_\mu]_{n_{c_\mu},2} & \cdots & [\mathsf{s}_\mu]_{n_{c_\mu},|S_\mu|-1} & [\mathsf{s}_\mu]_{n_{c_\mu},|S_\mu|} \end{matrix} \tag{25}$$

where $\mu = 1, 2$. With the existence of CNOT failures in the distillation circuits, the generalized syndromes in this array can be correlated between different rows. As an illustration, we study the following example first.

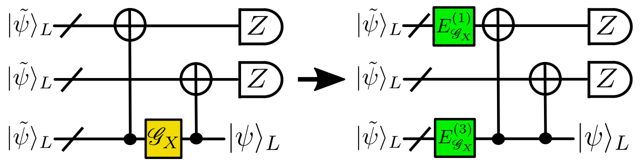

FIG. 7. An example of failures $\mathscr{G}_X$ associated with the CNOT gates on the third block. It is equivalent to have effective failures on the first and third blocks before a perfect distillation circuit.

Example 1. Consider the distillation circuit that uses the $[3, 1, 3]$ code to remove X failures. Let $\mathscr{G}$ be a failure set on the third block due to the transversal CNOTs from block 3 to block 1 as shown in the left circuit of Fig. 7. Since only X failures are supposed to be removed at this round, we only need to consider $\mathscr{G}_X$ and its support $\mathrm{QE}^{(3)}_{\mathscr{G}_X}$. Apparently, $\left|\mathrm{QE}^{(3)}_{\mathscr{G}_X}\right| = |\mathscr{G}_X|$, since $\mathscr{G}_X$ only has failures in one time step. The circuit is equivalent to a perfect distillation circuit with preceding failures $E^{(1)}_{\mathscr{G}_X} \cup E^{(3)}_{\mathscr{G}_X}$ on the first and third blocks. Since the circuit consists only of transversal gates, we have

$$\mathcal{I}\left(\mathrm{Q}^{(1)}_{E_{\mathscr{G}_X}}\right) = \mathcal{I}\left(\mathrm{Q}^{(3)}_{E_{\mathscr{G}_X}}\right) = \mathcal{I}\left(\mathrm{QE}^{(3)}_{\mathscr{G}_X}\right). \tag{26}$$

Thus, $\mathcal{I}\left(S^{(3)}_\mu|_{E_\mathscr{G}}\right) = \mathcal{I}\left(S^{(1)}_\mu|_{E_\mathscr{G}}\right) \equiv \mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}}$. For those syndrome columns with index $i \in \mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}}$, $[\mathsf{s}_\mu]_{1,i} = [\mathsf{s}_\mu]_{3,i} = 1$. Since the $[3, 1, 3]$ code can correct only one error, it will mistakenly decode syndrome column $i \in \mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}}$ and diagnose a flip in $[\mathsf{s}_\mu]_{2,i}$, leaving $E^{(3)}_{\mathscr{G}_X}$ on block 3 uncorrected after distillation.

Since the failures occur independently on the qubits, we are left with an X error of weight $w = \left|\mathrm{QE}^{(3)}_{\mathscr{G}_X}\right|$ with probability $p^{|\mathscr{G}|}$. By Def. 1, this is an uncorrelated error since $|\mathscr{G}| \geq |\mathscr{G}_X| = \left|\mathrm{QE}^{(3)}_{\mathscr{G}_X}\right|$, so this is fine. This fact suggests that certain types of CNOT failures should be harmless, and only those output qubits in the support of the faulty CNOTs would potentially be affected.

According to this observation, we can show that there are a large number of gate failures that are harmless as in Example 1. We have following result:

Lemma 1. *Consider a single round of the distillation protocol and, without loss of generality, suppose X errors are to be removed. Assume all the input ancilla states are error free. Suppose a gate failure set $\mathscr{G}$ occurs during distillation. If for any j, j' such that $QE^{(j)}_{\mathscr{G}_X}, QE^{(j')}_{\mathscr{G}_X} \neq \varnothing$ and*

$$\mathcal{I}\left(QE^{(j)}_{\mathscr{G}_X}\right) = \mathcal{I}\left(QE^{(j')}_{\mathscr{G}_X}\right), \tag{27}$$

then for each output block, it is either free of X errors or contains an X error of weight $w = \left|\mathcal{I}_{Q\,|\,\mathscr{G}_X}\right|$, where $\mathcal{I}_{Q\,|\,\mathscr{G}_X}$ is the set $\mathcal{I}\left(QE^{(j)}_{\mathscr{G}_X}\right)$ for some j such that $QE^{(j')}_{\mathscr{G}_X} \neq \varnothing$.

Proof. Since nonempty $\mathcal{I}\left(\mathrm{QE}^{(j)}_{\mathscr{G}_X}\right) = \mathcal{I}\left(\mathrm{Q}^{(j)}_{E_{\mathscr{G}_X}}\right)$ are all identical, according to the transversal property of the distillation circuit, when propagating the error back before the distillation circuit, the $\mathcal{I}\left(S^{(j)}_\mu\,|\,_{E_{\mathscr{G}_X}}\right)$ are identical for such a $S^{(j)}_\mu\,|\,_{E_{\mathscr{G}_X}} \neq \varnothing$. Let $\mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}} = \mathcal{I}\left(S^{(j)}_\mu\,|\,_{E_{\mathscr{G}_X}}\right)$ for such j. For any output block $r_{c_\mu} + 1 \leq j \leq n_{c_\mu}$, the estimated syndrome is that either $[\widetilde{\mathsf{s}}_\mu]_{j,i} = [\mathsf{s}_\mu]_{j,i}, \forall i \in \mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}}$, or $[\widetilde{\mathsf{s}}_\mu]_{j,i} \neq [\mathsf{s}_\mu]_{j,i}$, $\forall i \in \mathcal{I}_{S_\mu\,|\,E_{\mathscr{G}_X}}$. That is, either all the syndromes related to $\mathrm{Q}^{(j)}_{E_{\mathscr{G}_X}}$ are correctly estimated, or they all are correctly estimated, Consequently, after quantum error correction, the qubits with index in $\mathcal{I}_{\mathrm{Q}\,|\,\mathscr{G}_X}$ for each output block are either all free of error, or all left with X errors of weight $\left|\mathcal{I}_{\mathrm{Q}\,|\,\mathscr{G}_X}\right|$. □

Lemma 1 implies that if gate failures satisfying Eq. (27) are the only possible type of failures, then the output blocks are qualified, since the probability of weight $w \leq |\mathscr{G}|$ errors remaining in the output blocks will be $O(p^{|\mathscr{G}|})$. Such a set of gate failures is said to be *benign*. (Note that Example 1 is a special case, in that only the third block satisfies $\mathrm{QE}^{(3)}_{\mathscr{G}_X} \neq \varnothing$.)

2. Noisy preparation and distillation circuits

Now we consider the general case where the preparation circuit is also noisy. After encoding, errors of arbitrary weight can exist with probability $O(p)$, and the syndrome flips are independent among different blocks. However, there are additional correlated syndrome flips across different blocks due to the faulty CNOTs in the distillation circuit, which may compromise the validity of syndrome estimation. This is summarized in the following no-go theorem, which says that the previous distillation protocol cannot handle the situation where the distillation circuit is noisy.

Theorem 1 (No go). *Consider a single round of the distillation protocol by an $[n_{c_\mu}, k_{c_\mu}, d_{c_\mu}]$ classical code $\mathcal{C}_{c_\mu}$ with parity check matrix $\left[\mathsf{I}_{n_{c_\mu}} | \mathsf{A}_{c_\mu}\right]$. Denote the number of 1s in the jth column of A_{c_μ} by m_j and let $M = \max\{m_j\}$. If the distillation circuit is noisy with failure rate p, then the probability of an error of arbitrary weight w remaining on the output blocks will be $O(p^2)$ if $t_{c_\mu} \le M$, and $O(p^{t_{c_\mu}-M+2})$ if $t_{c_\mu} > M$, where $t_{c_\mu} = \lfloor \frac{d_{c_\mu}-1}{2} \rfloor$.*

Proof. Without loss of generality, we consider the round of distillation that is supposed to remove the X errors. Suppose we have a failure set $\mathscr{G}$, which contains a single CNOT failure in the distillation circuit involving block j_1. The support of $\mathscr{G}_X$ is $\mathrm{QE}_{\mathscr{G}_X}^{(j_1)}$. After the failure propagates to the beginning of the distillation circuit, it will introduce failures in a certain number of blocks. In the worst case, M blocks will be affected. Denoted this set of M blocks by $\mathcal{B}_{\mathscr{G}_X}$. Since the distillation circuit gates are all transversal, $\mathscr{G}_X$ is benign and $\mathcal{I}\left(S_\mu^{(j)}|_{E_{\mathscr{G}_X}}\right)$ are all identical to $\mathcal{I}\left(S_\mu^{(j_1)}|_{E_{\mathscr{G}_X}}\right)$ for $j \in \mathcal{B}_{\mathscr{G}_X}$ (denote it by $\mathcal{I}_{S_\mu \,|\, E_{\mathscr{G}_X}}$). Overall, $[\mathsf{s}_\mu]_{j,i}$ will be flipped to 1 for $j \in \mathcal{B}_{\mathscr{G}_X}$ and $i \in \mathcal{I}_{S_\mu \,|\, E_{\mathscr{G}_X}}$.

If $M \ge t_{c_\mu}$, then consider an additional single failure $\mathscr{R}$ that occurs in the previous stage on a block $j_2 \notin \mathcal{B}_{\mathscr{G}_X}$ such that for some $i_1 \in \mathcal{I}_{S_\mu \,|\, E_{\mathscr{G}_X}}$, $[\mathsf{s}_\mu]_{j_2,i_1} = 1$, and for all other $i \in \mathcal{I}_{S_\mu \,|\, E_{\mathscr{G}_X}}$, $[\mathsf{s}_\mu]_{j_2,i_1} = 0$. Thus, $[\widetilde{\mathsf{s}}_\mu]_{,:i_1} \neq [\mathsf{s}_\mu]_{,:i_1}$ because $[\mathsf{s}_\mu]_{,:i_1}$ has more than t_{c_μ} 1s. But all the other columns of syndromes are still correctly estimated. In the worst case, for a specific output block $r_{c_\mu} + 1 \le j_3 \le n_{c_\mu}$, syndromes for all stabilizers are correctly estimated except the i_1th one: In general, the decoded error $\widetilde{E}_X^{(j_3)}$ based on $[\widetilde{\mathsf{s}}_\mu]_{j_3,:}$ can be quite different from the actual X error on the jth block, $E_X^{(j_3)} = E_{\mathscr{G}_X}^{(j_3)} E_{\mathscr{R}_X}^{(j_3)}$. As a result, the residual error $\widetilde{E}_X^{(j_3)} E_X^{(j_3)}$ can have any weight. Thus, there is a probability $O(p^2)$ that an error of any weight w remains in the output blocks, since $\mathscr{G}$ and $\mathscr{R}$ are independent single failures.

For $M < t_{c_\mu}$, if $t_{c_\mu} - M + 1$ failures occur in $M - t_{c_\mu} + 1$ separated blocks outside $\mathcal{B}_{\mathscr{G}_X}$ in the previous stages, the situation is similar to the case where $M \ge t_{c_\mu}$, and this occurs with probability $O(p^{t_{c_\mu}-M+2})$. □

B. Fault-tolerant ancilla state distillation

In this subsection we show how calculating additional syndrome bits [37–39] can lead to the postselection of compatible error syndromes in the distillation protocol. As we will see, one can get round Theorem 1 by such a postselection, so that the resulting distillation protocol is fully fault-tolerant.

Consider the two groups generated by the stabilizer generators S_1, S_2 of the first and second rounds of distillation:

$$\mathcal{SC}_\mu = \langle S_\mu \rangle, \qquad \mu = 1, 2. \tag{28}$$

Definition 2. *For $\mu = 1, 2$, the eigenvalues associated with* all *the elements of $\mathcal{S}_\mu$ is called the* complete syndrome *for group $\mathcal{SC}_\mu$, and is denoted by $[\mathsf{sc}_\mu]_{j,i}$ for the ith syndrome bit in the jth block.*

Note that sc_μ can be estimated from the bitwise measurements of the check blocks. Denote the estimated complete syndrome array by $\widetilde{\mathsf{sc}}_\mu$. As before, let $\mathcal{SC}_\mu^{(j)}|_{E_{\mathscr{G}}}$ be the subset of stabilizers in $\mathcal{SC}_\mu^{(j)}$ that anticommute with $E_{\mathscr{G}}$. Using the distillation protocol proposed in the previous section, one can estimate the eigenvalue of any element in $\mathcal{SC}_\mu$ (syndromes) with only additional classical computation and no additional physical operations.

Now consider three elements $\mathrm{SC}_{\mu,1} \in S_\mu$, $\mathrm{SC}_{\mu,2} \in S_\mu$ and $\mathrm{SC}_{\mu,3} = \mathrm{SC}_{\mu,1}\mathrm{SC}_{\mu,2} \in \mathcal{SC}_\mu \backslash S_\mu$. Denote the corresponding estimated generalized syndrome bits on a specific output block j by $[\widetilde{\mathsf{sc}}]_{j,1}$, $[\widetilde{\mathsf{sc}}]_{j,2}$ and $[\widetilde{\mathsf{sc}}]_{j,3}$, which are supposed to satisfy $[\widetilde{\mathsf{sc}}]_{j,3} = [\widetilde{\mathsf{sc}}]_{j,1} + [\widetilde{\mathsf{sc}}]_{j,2}$. In such case, the estimated syndrome bits are said to be *compatible*. In general, although the values of $[\widetilde{\mathsf{sc}}]_{j,3}$ and $[\widetilde{\mathsf{sc}}]_{j,1} + [\widetilde{\mathsf{sc}}]_{j,2}$ are correlated, they are not always the same when failures occur. If $[\widetilde{\mathsf{sc}}]_{j,3} \neq [\widetilde{\mathsf{sc}}]_{j,1} + [\widetilde{\mathsf{sc}}]_{j,2}$, it strongly implies that the estimated syndrome set is problematic and the corresponding ancilla block cannot be used. This method can be extended to check any number of products of elements in S_μ.

We therefore propose a modified distillation protocol by estimating additional syndrome bits and checking whether the estimated syndrome bits are compatible with each other. If they are not, we discard the distilled ancillas and try again.

In this paper, we use a classical error-detecting code to define a suitable set of additional stabilizers in $\mathcal{SC}_\mu \backslash S_\mu$.

Consider an $[n_{d_\mu}, k_{d_\mu} = |S_\mu|]$ classical code $\mathcal{C}_{d_\mu}$ with parity-check matrix $\mathsf{H}_{d_\mu} = [\mathsf{I}_{n_{d_\mu}-|S_\mu|} \mid \mathsf{A}_{d_\mu}]$, where A_{d_μ} is an $(n_{d_\mu} - |S_\mu|) \times |S_\mu|$ binary matrix. We define a set of additional stabilizers S'_μ where

$$S'_{\mu_j} = \prod_{i=1}^{|S_\mu|} S_{\mu_i}^{[\mathsf{A}_{d_\mu}]_{j,i}}, \qquad j = 1, \ldots, n_{d_\mu} - |S_\mu|, \tag{29}$$

where S'_{μ_j} is the jth element of S'_μ and S_{μ_i} us the ith element of S_μ. S'_μ and S_μ can be combined to form an extended set of stabilizer $SE_\mu = S'_\mu \bigcup S_\mu$, whose eigenvalues need to be estimated. The corresponding extended syndrome array of SE_μ for all blocks can be defined as

$$\mathsf{se}_\mu = [\mathsf{s}'_\mu | \, \mathsf{s}_\mu]. \tag{30}$$

After classical decoding, each output block j has an estimated syndrome vector $[\widetilde{\mathsf{se}}_\mu]_{j,:}$. By the definition of S'_μ, $[\widetilde{\mathsf{se}}_\mu]_{j,:}$ is compatible only if

$$\mathsf{H}_{d_\mu} [\widetilde{\mathsf{se}}_\mu]_{j,:}^T = 0. \tag{31}$$

If the block does not pass this test, its estimated syndrome bits are likely to be chaotic. Since for each j, the values of

$[\widetilde{\mathsf{se}}_\mu]_{j,i}$ are strongly correlated between different i, we can not do error correction using $\mathcal{C}_{d_\mu}$ but instead simply discard the outputs. Such process is called *postselection*: we accept the output ancillas if Eq. (31) holds, and we do an error correction procedure based on $[\widetilde{\mathsf{s}}_\mu]$ for the remaining blocks as before. This scheme is fault-tolerant only if the high-weight errors remaining on the output blocks are greatly suppressed by postselection, and the ancillas that pass the postselection are qualified. In addition, we want that the rejection rate not to be too high. In the next subsection, we will address these issues in detail.

C. Fault-tolerance of the protocol

In this section, we study the conditions such that the proposed protocol in the previous subsection is fault-tolerant. A glossary is provided in Appendix. B (Table I) to help the readers with the various symbols used in the text.

In the previous subsection, we use a classical error-detecting code to define additional syndrome bits and check whether they are compatible. In the ideal case, if the complete syndrome set is estimated and checked for compatibility, the postselection will work better than the nonideal case, since it has the most error information.

Definition 3. *Suppose the complete syndrome set of block j for $\mathcal{SC}_\mu^{(j)}$ is estimated ($\mu = 1, 2$). For any $SC_{\mu,i_l} \in \langle S_\mu \rangle$ with a decomposition $SC_{\mu,i_1} \times \cdots \times SC_{\mu,i_k}$, where $SC_{\mu,i_j} \in S_\mu$, if the estimated syndrome bits satisfy*

$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_l} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_1} + \cdots + [\widetilde{\mathsf{sc}}_\mu]_{j,i_k}, \tag{32}$$

then the estimated complete syndrome set is said to be valid *for block j.*

Remark 1. Even if the estimated complete syndrome set is valid, it is not necessarily correct; instead, it simply means that the estimated syndromes are compatible with each other and thus more trustworthy.

Definition 4. *A postselection is said to be* ideal *if those output blocks with valid estimated complete syndrome sets are kept and the rest are discarded in a round of distillation.*

As we have seen in the proof of Lemma 1, certain syndrome patterns caused by a benign failure set are harmless. We can define such syndrome patterns as follows.

Definition 5. *Consider the complete syndrome set for all the ancilla blocks. If the values of the syndrome bits for each block are identical, then this complete syndrome array is said to be* good*. The subset of blocks containing these non-zero syndrome bits are denoted as $\mathcal{B}_f$.*

An example of a good syndrome pattern is shown in Fig. 8 for five blocks. One can directly have the following lemma from the observation in Sec. IV A:

Lemma 2. *A benign failure set causes a good syndrome pattern before distillation.*

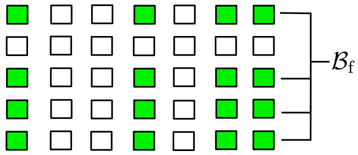

FIG. 8. (Color online) An illustration of a good syndrome pattern for syndrome array sc_μ, with $|S_\mu| = 3$. The green boxes represent 1-value syndromes and white ones represent 0-value syndromes.

Another useful property of a good syndrome pattern is as follows.

Lemma 3. *All the nonzero columns of the syndrome array of a good syndrome pattern are identical.*

Since a good syndrome pattern is harmless, we may expect that it will pass ideal postselection. Indeed, we have the following lemma:

Lemma 4. *If the complete syndrome pattern is good, then the estimated complete syndrome set is valid.*

Proof. If the number of blocks with syndrome bits flipped to 1 is no more than t_c, then the estimated complete syndrome sets are all correct, that is, $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$ for all j and i, and the estimated complete syndrome sets are automatically valid. Otherwise, the estimated syndromes can either be wrong or correct, depending on the classical decoding procedure.

Consider two columns i_1 and i_2 and block $j \in \mathcal{B}_\mathrm{f}$. Suppose $\mathrm{SC}_{\mu,i_3} = \mathrm{SC}_{\mu,i_1}\mathrm{SC}_{\mu,i_2} \in \langle S_\mu \rangle$. If $[\mathsf{sc}_\mu]_{j,i_1} = 1$ and $[\mathsf{sc}_\mu]_{j,i_2} = 0$, then $[\mathsf{sc}_\mu]_{j,i_3} = 1$. Then $[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_3}$ because $[\mathsf{sc}]_{:,i_1} = [\mathsf{sc}]_{:,i_3}$ when the syndrome pattern is good. Meanwhile, $[\mathsf{sc}_\mu]_{:,i_2}$ must be all 0 according to Lemma 3, and thus $[\widetilde{\mathsf{sc}}_\mu]_{j,i_2} = 0$. Then one obtains

$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} + [\widetilde{\mathsf{sc}}_\mu]_{j,i_2} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_3}.$$

The above equation holds for arbitrary values of $[\mathsf{sc}_\mu]_{j,i_1}$ and $[\mathsf{sc}_\mu]_{j,i_2}$, . Similarly, we can see that

$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_l} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_1} + \cdots + [\widetilde{\mathsf{sc}}_\mu]_{j,i_k}$$

holds in general for $\mathrm{SC}_{\mu,i_l} = \mathrm{SC}_{\mu,i_1} \cdots \mathrm{SC}_{\mu,i_k} \in \langle S_\mu \rangle$. □

A natural question arises: can we mainly rely on such a postselection to reduce high weight errors? In other words, can we just use some classical codes of very small distance for $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$, say distance three, to roughly estimate the general syndrome bits, and then take error detection codes $\mathcal{C}_{d_1}$ and $\mathcal{C}_{d_2}$ of large distance to remove the blocks with incompatible syndrome bits? Then after quantum error correction, would the output ancillas be qualified?

To answer these questions, first consider a perfect distillation circuit and check how well the postselection mechanism can perform. Can we suppress the error rate to order higher than $t_c + 1$? The answer is negative because of the correlation among the estimated general syndrome bits after classical decoding. Here, we show the limitation of the postselection mechanism alone:

Theorem 2. *For a single round of the distillation process, if the classical error-correcting code $\mathcal{C}_{c_\mu}$ can correct $t_{c_\mu} < t-1$ errors, where the underlying quantum code can correct t errors, then the probability of any error of weight $w > t$ on the output is at most $O(p^{t_{c_\mu}+1})$ even for ideal postselection.*

Proof. Without loss of generality, let us assume this round of distillation is to remove X errors. Suppose the failure set $\mathscr{G}$ occurs in the preparation stage and the distillation circuit is perfect. Assume the support of $\mathscr{G}_X$ is over $t_{c_\mu}+1$ blocks. Denote this set of $t_{c_\mu}+1$ blocks by $\mathcal{B}_{\mathscr{G}_X}$. Suppose that $\mathcal{I}\left(\mathrm{Q}^{(j)}_{E_{\mathscr{G}_X}}\right)$ is the same for $j \in \mathcal{B}_{\mathscr{G}_X}$:

$$\mathcal{I}\left(\mathrm{Q}^{(j)}_{E_{\mathscr{G}_X}}\right) \triangleq \mathcal{I}_{\mathrm{Q}\,|\,E_{\mathscr{G}_X}}, \qquad j \in \mathcal{B}_{\mathscr{G}_X}. \tag{33}$$

This means that each block in $\mathcal{B}_{\mathscr{G}_X}$ has exactly the same effective X error $E^{(j)}_{\mathscr{G}_X}$ before distillation. Obviously, $\mathrm{Q}^{(j)}_{E_{\mathscr{G}_X}}$ for $j \in \mathcal{B}_{\mathscr{G}_X}$ shares the same index set, denoted by $\mathcal{I}_{\mathrm{Q}\,|\,E_{\mathscr{G}_X}}$. Thus, $\mathcal{I}\left(\mathcal{SC}^{(j)}_\mu|_{E_{\mathscr{G}_X}}\right)$ is also the same for all $j \in \mathcal{B}_{\mathscr{G}_X}$:

$$\mathcal{I}\left(\mathcal{SC}^{(j)}_\mu|_{E_{\mathscr{G}_X}}\right) \triangleq \mathcal{I}_{\mathcal{SC}_\mu\,|\,E_{\mathscr{G}_X}}, \qquad j \in \mathcal{B}_{\mathscr{G}_X}. \tag{34}$$

Obviously, $[\mathsf{sc}_\mu]_{j,i} = 1$ for $\forall i \in \mathcal{I}_{\mathcal{SC}_\mu\,|\,E_{\mathscr{G}_X}}$, $\forall j \in \mathcal{B}_{\mathscr{G}_X}$, otherwise $[\mathsf{sc}_\mu]_{j,i} = 0$. This syndrome pattern is good, and according to Lemma 4 the estimated complete syndrome set will be valid for all blocks. Thus, any output block will pass ideal postselection. On the other hand, since $|\mathcal{B}_{\mathscr{G}_X}| = t_{c_\mu}+1 > t_{c_\mu}$, $[\widetilde{\mathsf{sc}}_\mu]_{:,i} \neq [\mathsf{sc}_\mu]_{:,i}$ for $i \in \mathcal{I}_{\mathcal{SC}_\mu\,|\,E_{\mathscr{G}_X}}$. But one still has $[\widetilde{\mathsf{sc}}_\mu]_{:,i_1} = [\widetilde{\mathsf{sc}}_\mu]_{:,i_2}$ for any $i_1, i_2 \in \mathcal{I}_{\mathcal{SC}_\mu\,|\,E_{\mathscr{G}_X}}$. Therefore, for any specific output block $r_{c_\mu}+1 \le j \le n_{c_\mu}$ either $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$ or $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$. In the worst case, $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, and X errors are left on qubits in $\mathcal{I}_{\mathrm{Q}\,|\,E_{\mathscr{G}_X}}$ after quantum error correction. Note that the failure set $\mathscr{G}$ occurs with probability $O(p^{|\mathscr{G}|}) \le O(p^{t_{c_\mu}+1})$ and $\left|\mathcal{I}_{\mathrm{Q}\,|\,E_{\mathscr{G}_X}}\right|$ can be arbitrary. So there always exists a $\mathscr{G}$ that leaves an X error of weight $w = \left|\mathcal{I}_{\mathrm{Q}\,|\,E_{\mathscr{G}_X}}\right| > t$ on output block j with probability $O(p^{t_{c_\mu}+1})$, after ideal postselection and quantum error correction. □

Complementary to Lemma 4, one may infer that the pattern of a complete syndrome array is close to a good one if any of the output blocks passes ideal postselection. For simplicity, we define the concept of ϵ-sparse for a quantum code $\mathcal{Q}$:

Definition 6. *Consider a quantum code $\mathcal{Q}$. Let $\mathbb{S}_\cap$ ($\mathbb{S}_{\not\cap}$) be the set of complete syndrome pairs $(\mathsf{sc1}_\mu, \mathsf{sc2}_\mu)$, with $\mathsf{sc1}_\mu \neq \mathsf{sc2}_\mu$, such that they correspond to two errors $E_{1\mu}, E_{2\mu}$ whose supports have an nonempty (empty) intersection. If $|\mathbb{S}_\cap|/|\mathbb{S}_{\not\cap}| < \epsilon$, $\mathcal{Q}$ is called an ϵ-*sparse code.

Quantum codes such as surface codes and quantum LDPC codes have low ϵ when the code length is large.

Lemma 5. *Suppose $\mathcal{Q}$ is ϵ-sparse. If the estimated complete syndrome set is valid for a specific output block, then the corresponding syndrome array for all input blocks is, with probability at least $1 - O(\epsilon)$, either a good syndrome pattern or close to a good syndrome pattern.*

Proof. Without loss of generality, assume output block j has a valid complete syndrome set and its X errors are removed. The first observation is that if $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$, then it is likely the whole column of syndrome bits $[\widetilde{\mathsf{sc}}_\mu]_{:,i}$ was correctly decoded by $\mathcal{C}_{c_\mu}$ ($[\widetilde{\mathsf{sc}}_\mu]_{:,i} = [\mathsf{sc}_\mu]_{:,i}$), and thus the number of 1s in $[\mathsf{sc}_\mu]_{j,i}$ is fewer than t_{c_μ}. Similarly, if $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}$, then the column of syndrome bits $[\widetilde{\mathsf{sc}}_\mu]_{:,i}$ is incorrectly estimated. (Note that it is possible that $[\mathsf{sc}_\mu]_{:,i} \neq [\widetilde{\mathsf{sc}}_\mu]_{:,i}$, but the estimate on the jth block is still correct: $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$. See Fig. 10 for an example. We will address this case later.)

Since $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ is valid, Eq. (32) must be satisfied, that is, for $\mathrm{SC}_{\mu,i_3} = \mathrm{SC}_{\mu,i_1}\mathrm{SC}_{\mu,i_2} \in \langle S_\mu \rangle$, $[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} + [\widetilde{\mathsf{sc}}_\mu]_{j,i_2} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_3}$ must hold. If $[\widetilde{\mathsf{sc}}_\mu]_{:,i_1} \neq [\mathsf{sc}_\mu]_{:,i_1}$ and $[\widetilde{\mathsf{sc}}_\mu]_{:,i_2} \neq [\mathsf{sc}_\mu]_{:,i_2}$, then $[\widetilde{\mathsf{sc}}_\mu]_{j,i_1}$ and $[\widetilde{\mathsf{sc}}_\mu]_{j,i_2}$ are likely to be incorrectly estimated. It then is necessary that $[\widetilde{\mathsf{sc}}_\mu]_{j,i_3} = [\mathsf{sc}_\mu]_{j,i_3}$. Consequently, it is likely that $[\widetilde{\mathsf{sc}}_\mu]_{:,i_3} = [\mathsf{sc}_\mu]_{:,i_3}$. Similarly, if $[\widetilde{\mathsf{sc}}_\mu]_{:,i_1} \neq [\mathsf{sc}_\mu]_{:,i_1}$ but $[\widetilde{\mathsf{sc}}_\mu]_{:,i_2} = [\mathsf{sc}_\mu]_{:,i_2}$, then it is likely $[\widetilde{\mathsf{sc}}_\mu]_{:,i_3} \neq [\mathsf{sc}_\mu]_{:,i_3}$.

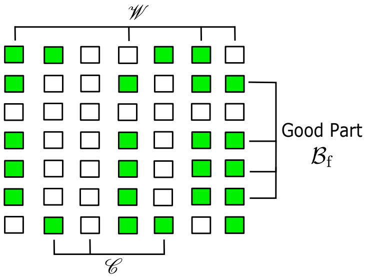

FIG. 9. (Color online) An illustration of an approximately good syndrome pattern for syndrome array sc_μ. Here, we assume that $\mathcal{C}_{c_\mu}$ can correct two errors and $|S_\mu| = 3$. The green boxes represent syndrome bit 1s and the white ones represent syndrome bit 0s. The good part of this pattern, $\mathcal{B}_\mathrm{f}$, represents the maximum subset of rows with identical syndromes.

We can generalize this observation to more than three columns of syndromes. First, denote the set of columns such that $[\widetilde{\mathsf{sc}}_\mu]_{:,i} \neq [\mathsf{sc}_\mu]_{:,i}$ as $\mathscr{W}$. Take any $2x$ columns $[\mathsf{sc}_\mu]_{:.i_1}, \ldots, [\mathsf{sc}_\mu]_{:.i_{2x}} \in \mathscr{W}$, and suppose $\mathrm{SC}_{\mu,i_l} = \mathrm{SC}_{\mu,i_1} \times \cdots \times \mathrm{SC}_{\mu,i_{2x}} \in \langle S_\mu \rangle$. Since

$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} +, \ldots, +[\widetilde{\mathsf{sc}}_\mu]_{j,i_{2x}} = [\widetilde{\mathsf{sc}}_\mu]_{j,i_l},$$

$[\widetilde{\mathsf{sc}}_\mu]_{j,i_l} = [\mathsf{sc}_\mu]_{j,i_l}$ from the validity of output block j. Thus, the column of syndrome $[\mathsf{sc}_\mu]_{:,i_l}$ is likely equal to $[\widetilde{\mathsf{sc}}_\mu]_{:,i_l}$ and contains less than t_c 1s. Consequently, the columns in $\mathscr{W}$ have a large amount of overlap in syndrome bits that are 1s. Denote the set of columns such that $[\widetilde{\mathsf{sc}}_\mu]_{:,i} = [\mathsf{sc}_\mu]_{:,i}$ as $\mathscr{C}$. Consider any $[\mathsf{sc}_\mu]_{:,i} \in \mathscr{W}$ and $[\mathsf{sc}_\mu]_{:,i'} \in \mathscr{C}$. Suppose $\mathrm{SC}_i\mathrm{SC}_{i'} = \mathrm{SC}_{i''} \in \langle S_\mu \rangle$. Since $[\widetilde{\mathsf{sc}}_\mu]_{j,i} + [\widetilde{\mathsf{sc}}_\mu]_{j,i'} = [\widetilde{\mathsf{sc}}_\mu]_{j,i''}$, it is likely that $[\mathsf{sc}_\mu]_{:,i''} \in \mathscr{W}$. Since we know that $[\mathsf{sc}_\mu]_{:,i}$ and $[\mathsf{sc}_\mu]_{:,i''}$ have large overlaps, it also implies that for any $[\mathsf{sc}_\mu]_{:,i'} \in \mathscr{C}$, the number of 1s it contains is likely to be small, which is consistent with our previous observation. Such syndrome array parttern is called an *approximately good* syndrome pattern. Figure 9 illustrates such a pattern. The overlaps of different columns in $\mathscr{W}$ are called the good part of the pattern. So the syndrome pattern is likely to be close to a good

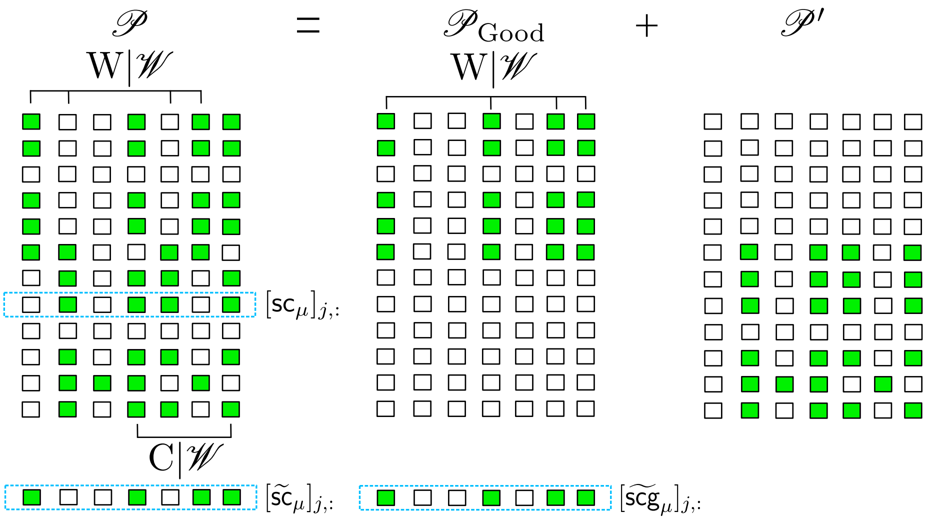
$\mathscr{P}$
=
$\mathscr{P}_{\text{Good}}$
+
$\mathscr{P}'$
W|$\mathscr{W}$
W|$\mathscr{W}$
$[\mathsf{sc}_\mu]_{j,:}$
C|$\mathscr{W}$
$[\widetilde{\mathsf{sc}}_\mu]_{j,:}$
$[\widetilde{\mathsf{scg}}_\mu]_{j,:}$

FIG. 10. (Color online) An illustration of a pattern $\mathscr{P}$ that is not close to good pattern, but has valid estimated complete syndrome set on output block j after classical decoding. Here, we assume that $\mathcal{C}_{c_\mu}$ can correct two errors, and $|S_\mu| = 3$. The green boxes represent 1-value syndromes and white ones represent 0-value syndromes. The complete syndromes of output block j are circled by a dashed line and its estimated syndrome $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ is shown at the bottom of the left figure. $\mathscr{P}$ can always be decomposed into a good syndrome pattern $\mathscr{P}_{\text{Good}}$ and another pattern $\mathscr{P}'$, such that $\mathscr{P}_{\text{Good}}$ generates exactly the same $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ after classical decoding. In this example, $\mathscr{P}_{\text{Good}}$ is constructed as follows: 1. each column $[\mathsf{scg}]_{:,i}$ such that $[\widetilde{\mathsf{sc}}]_{j,i} = 0$ is set to be an all-zero syndrome column; 2. each column $[\mathsf{scg}]_{:,i}$ such that $[\widetilde{\mathsf{sc}}]_{j,i} = 1$ is set to be the first column of $\mathscr{P}$ (which gives $[\widetilde{\mathsf{sc}}]_{j,i} = 1$ after classical decoding.).

.

pattern when any specific output block has a valid estimated complete syndrome set.

Now we consider those syndrome patterns whose columns in $\mathscr{W}$ do not have a large amount of overlap, but still generate a valid estimated complete syndrome set for output block j. In other words, the complete syndrome pattern is not good or approximately good. An example of such a pattern is shown in Fig. 10. One can further separate the column set of $\mathscr{W}$ into two types: W|$\mathscr{W}$ and C|$\mathscr{W}$. For $[\mathsf{sc}_\mu]_{:,i} \in$ W|$\mathscr{W}$, one has $[\widetilde{\mathsf{sc}}_\mu]_{:,i} \neq [\mathsf{sc}_\mu]_{:,i}$ and $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}$; while for $[\mathsf{sc}_\mu]_{:,i} \in$ C|$\mathscr{W}$, one has $[\widetilde{\mathsf{sc}}_\mu]_{:,i} \neq [\mathsf{sc}_\mu]_{:,i}$ but $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$.

It is always possible to find a good pattern $\mathscr{P}_{\text{Good}}$ whose complete syndrome array scg_μ generates exactly the same estimated complete syndromes for block j as sc_μ ($[\widetilde{\mathsf{scg}}_\mu]_{j,:} = [\widetilde{\mathsf{sc}}_\mu]_{j,:}$) after classical decoding. This $\mathscr{P}_{\text{Good}}$ can be constructed as follows: If $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = 0$, then $[\mathsf{scg}_\mu]_{:,i}$ is set to be an all-zero column; if $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = 1$, then choose an arbitrary column $[\mathsf{sc}_\mu]_{:,i'}$ in $\mathscr{P}$ such that $[\widetilde{\mathsf{sc}}_\mu]_{j,i'} = 1$, and then set $[\mathsf{scg}_\mu]_{:,i} = [\mathsf{sc}_\mu]_{:,i'}$. We can then define another pattern $\mathscr{P}'$ as the difference between $\mathscr{P}$ and $\mathscr{P}_{\text{Good}}$, whose complete syndrome set array is denoted by sc'_μ. From this construction, since $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ is valid, scg_μ (and also sc'_μ) is legal, in the sense that its syndromes are compatible in each block. An example of this construction is shown in Fig. 10. Note that both $\mathscr{P}$ and $\mathscr{P}_{\text{Good}}$ give the same $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ by the same classical decoding procedure.

Now, we show that the probability of pattern $\mathscr{P}$ is much smaller than that of $\mathscr{P}_{\text{Good}}$. Suppose the failure sets $\mathscr{G}$, $\mathscr{G}_{\text{Good}}$ and $\mathscr{G}'$ induce $\mathscr{P}$, $\mathscr{P}_{\text{Good}}$ and $\mathscr{P}'$, respectively. For each j' such that $[\mathsf{sc}_\mu]_{j',:} \neq 0$, we have $[\mathsf{scg}_\mu]_{j',:} \neq [\mathsf{sc}'_\mu]_{j',:}$, if $[\mathsf{sc}_\mu]_{j,:}$ is not all zero (see Fig. 10, for example). Then, for each j', $\Pr\left\{\mathrm{Q}^{(j')}_{E_{\mathscr{G}_{X\text{Good}}}} \cap \mathrm{Q}^{(j')}_{E_{\mathscr{G}'_X}} \neq \varnothing\right\} < \epsilon$ by assumption. (This probability should be much smaller, since the non-trivial contribution of $|\mathbb{S}_\cap|/|\mathbb{S}_{\not\cap}|$ comes generally from high weight errors in Def. 6, which are less likely to happen.) This suggests that for every time step τ of the procedure, $\Pr\left\{\mathrm{Q}^{(j)}_{\mathscr{G}\mathscr{M}_{X\text{Good}}(\tau)} \cap \mathrm{Q}^{(j)}_{\mathscr{G}\mathscr{M}'_X(\tau)} \neq \varnothing\right\} < \epsilon$. Here, $\mathscr{G}\mathscr{M}_{X\text{Good}}$ and $\mathscr{G}\mathscr{M}'_X$ are the failure sets which have the highest probability among those failure sets $\mathscr{G}_{X\text{Good}}$ and $\mathscr{G}'_X$ causing $\mathscr{P}_{\text{Good}}$ and $\mathscr{P}'$. As a result, $\mathscr{G}\mathscr{M}_{X\text{Good}}$ and $\mathscr{G}\mathscr{M}'_X$ are independent with probability at least $1 - O(\epsilon)$. We then have

$$\max_{\mathscr{G}} \Pr(\mathscr{P}) \leq \max_{\mathscr{G}_{\text{Good}},\mathscr{G}'} \Pr(\mathscr{P}_{\text{Good}})\Pr(\mathscr{P}'), \tag{35}$$

which implies that there exists at least one decomposition of $\mathscr{P}$ such that $\Pr(\mathscr{P}) \ll \Pr(\mathscr{P}_{\text{Good}})$ for all $\mathscr{G}$ causing $\mathscr{P}$. Thus, one can always find another failure set that generates a good syndrome pattern which causes the same valid complete estimated syndrome set $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ with a much higher probability.

Overall, the corresponding syndrome array for all input blocks is most likely to be either good syndrome pattern or close to good pattern if $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ is valid.

□

Theorem 2 states the limitation on suppression of the high weight errors by a postselection protocol alone, and suggests that t_{c_μ} must be no less than $t - 1$ to get qualified output ancillas. Moreover, we can prove that when this is true, an output block that passes ideal postselection is qualified, for sufficiently large block codes:

Theorem 3. *Suppose $\mathcal{Q}$ is an ϵ-sparse quantum code. Then if the estimated complete syndrome set for one output block is valid, and $t_{c_\mu} \geq t-1$, this output block is qualified with probability at least $1-O(\epsilon)$.*

Proof. Consider a round of distillation to remove X errors with output block j. Since $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$ can pass ideal postselection, it is valid. From Lemma 5, the syndrome pattern $\mathscr{P}$ is either good or approximately good with probability at least $1-O(\epsilon)$.

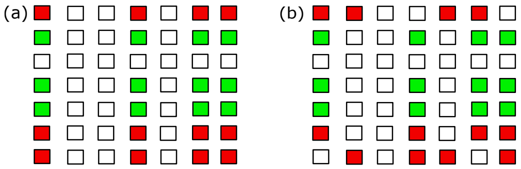

FIG. 11. (Color online) An illustration of an approximately good syndrome pattern for syndrome array sc_μ caused by $\mathscr{G}_X$ (green boxes) and $\mathscr{R}_X$ (red boxes) for good pattern (a) and approximately good pattern (b).

Consider first the case when $\mathscr{P}$ is good (see Fig. 11 (a) for an illustration). If the number of nonzero rows of sc, say $|\mathcal{B}_{\mathrm{f}}|$, is less or equal to t_c, then $[\widetilde{\mathsf{sc}}_\mu]_{j,:} = [\mathsf{sc}_\mu]_{j,:}$ after classical decoding and no X error is left after quantum error correction. If $|\mathcal{B}_{\mathrm{f}}| > t_{c_\mu}$, we assume $\mathscr{P}$ is caused by a combination of two failure sets: 1) a benign CNOT failure set $\mathscr{G}_X$ in the distillation circuit that flips certain rows of the syndrome array to 1 (which are shown as the green squares in Fig. 11 (a)); and 2) the remaining CNOT failures in the distillation circuit and the failures from the preparation stage which together form a failure set $\mathscr{R}_X$ that flips the rest of the nonzero rows of the syndrome array (which are shown as the red squares in Fig. 11 (a)). Denote by $\mathcal{B}_{\mathscr{G}_X}$ and $\mathcal{B}_{\mathscr{R}_X}$ the index sets of the rows of the syndrome array affected by $\mathscr{G}_X$ and $\mathscr{R}_X$, respectively. Consider the following two cases:

1. $\mathscr{G}_X \neq \varnothing$. For all $j' \in \mathcal{B}_{\mathscr{G}_X}$, one has $\mathrm{QE}^{(j')}_{\mathscr{G}_X} = \mathrm{Q}^{(j')}_{E_{\mathscr{G}_X}} \neq \varnothing$, since the $\mathscr{G}_X$ is benign. Since the distillation circuit is transversal, for all $j' \in \mathcal{B}_{\mathscr{G}_X}$, one has
$$\mathcal{I}\left(\mathcal{SC}^{(j')}_\mu|_{E_{\mathscr{G}_X}}\right) = \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}} = \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{R}_X}}, \tag{36}$$
where the second equality is because the syndrome pattern is good. Then, for output block j, either $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, or $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\widetilde{\mathsf{sc}}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$ after classical decoding. In the worst case, $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, which will leave an X error on output block j. The support is the same as $\mathcal{I}\left(\mathrm{QE}^{(j')}_{\mathscr{G}_X}\right)$. This case occurs with probability $O(p^{|\mathscr{G}_X|+|\mathscr{R}_X|})$. Since $w = \left|\mathcal{I}\left(\mathrm{QE}^{(j')}_{\mathscr{G}_X}\right)\right| \leq |\mathscr{G}_X|$, block j is qualified.

2. $\mathscr{G}_X = \varnothing$. Now all the colored squares in Fig. 11 (a) will be red, and this occurs with probability $O(p^{|\mathscr{R}_X|})$. In this case, an X error of arbitrary weight (depending on $\mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$) may remain in output block j. Since now $|\mathscr{R}_X| \geq |\mathcal{B}_{\mathscr{R}_X}| > t_{c_\mu} \geq t-1$, it is uncorrelated, and block j is still qualified.

Now we consider the case of an approximately good syndrome pattern $\mathscr{P}_A$ (see Fig. 11 (b) for an illustration). The argument is similar here. If the maximum number of nonzero rows in $[\mathsf{sc}_\mu]_{:,i}$ is less then t_{c_μ}, $[\widetilde{\mathsf{sc}}_\mu]_{j,:} = [\mathsf{sc}_\mu]_{j,:}$ after classical decoding, and it leaves no error in the output block after quantum error correction. If this is not the case, we can also decompose the failures that lead to $\mathscr{P}_A$ into two sets: a benign CNOT failure set $\mathscr{G}_X$ in the distillation circuit that flips certain identical rows of syndrome bits (the green squares in Fig. 11 (b)), and the remaining CNOT failures in the distillation circuit combined with the failures $\mathscr{R}_X$ in the preparation stage, which flip the syndrome bits of the rest of the nonzero rows (the red squares in Fig. 11 (b)). $\mathcal{B}_{\mathscr{G}_X}$ and $\mathcal{B}_{\mathscr{R}_X}$ can be defined in the same way as for good pattern, and $\mathcal{B}_{\mathscr{G}_X}$ lies in the good part $\mathcal{B}_{\mathrm{f}}$ of $\mathscr{P}_A$. If $\mathscr{G}_X = \varnothing$, the situation is similar to the case of a good syndrome pattern. An X error of arbitrary weight may remain in output block j with probability $O(p^{|\mathscr{R}_X|})$, where $|\mathscr{R}_X| \geq |\mathcal{B}_{\mathscr{R}_X}| > t_{c_\mu} \geq t-1$. When $\mathscr{G}_X \neq \varnothing$, for all $j' \in \mathcal{B}_{\mathscr{G}_X}$ one has $\mathrm{QE}^{(j')}_{\mathscr{G}_X} = \mathrm{Q}^{(j')}_{E_{\mathscr{G}_X}} \neq \varnothing$. Since the distillation circuit is transversal, for all $j' \in \mathcal{B}_{\mathscr{G}_X} \subset \mathcal{B}_{\mathrm{f}}$, one has
$$\mathcal{I}\left(\mathcal{SC}^{(j')}_\mu|_{E_{\mathscr{G}_X}}\right) = \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}. \tag{37}$$
Since $\mathscr{P}_A$ is an approximately good pattern, for $i \notin \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, $[\mathsf{sc}_\mu]_{:,i} \in \mathscr{C}$ and $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$ for output block j (see Fig. 9). For $i_1, i_2 \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$ and $\mathrm{SC}_{\mu,i_3} = \mathrm{SC}_{\mu,i_1}\mathrm{SC}_{\mu,i_2} \in \langle S_\mu \rangle$, $[\mathsf{sc}_\mu]_{:,i_3} = [\mathsf{sc}_\mu]_{:,i_1} + [\mathsf{sc}_\mu]_{:,i_2} \in \mathscr{C}$, so either
$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} = [\mathsf{sc}_\mu]_{j,i_1}, \quad [\widetilde{\mathsf{sc}}_\mu]_{j,i_2} = [\mathsf{sc}_\mu]_{j,i_2}, \tag{38}$$
or
$$[\widetilde{\mathsf{sc}}_\mu]_{j,i_1} \neq [\mathsf{sc}_\mu]_{j,i_1}, \quad [\widetilde{\mathsf{sc}}_\mu]_{j,i_2} \neq [\mathsf{sc}_\mu]_{j,i_2}, \tag{39}$$
because of the validity of $[\widetilde{\mathsf{sc}}_\mu]_{j,:}$. As a result, either $[\widetilde{\mathsf{sc}}_\mu]_{j,i} = [\mathsf{sc}_\mu]_{j,i}$ $\forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, or $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\widetilde{\mathsf{sc}}_\mu]_{j,i}$ $\forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$ after classical decoding. In the worst case, $[\widetilde{\mathsf{sc}}_\mu]_{j,i} \neq [\mathsf{sc}_\mu]_{j,i}, \forall i \in \mathcal{I}_{\mathcal{SC}_\mu|E_{\mathscr{G}_X}}$, which will leave X errors on the qubits corresponding to the index set $\mathcal{I}\left(\mathrm{QE}^{(j')}_{\mathscr{G}_X}\right)$ on block j with probability $O(p^{|\mathscr{G}_X|+|\mathscr{R}_X|})$. Since $w = \left|\mathcal{I}\left(\mathrm{QE}^{(j')}_{\mathscr{G}_X}\right)\right| \leq |\mathscr{G}_X|$, block j is qualified.

In summary, output block j is qualified with probability $1-O(\epsilon)$ if it can pass ideal postselection. □

Combining the above two theorems, we have the following main theorem of this paper:

Theorem 4. *Suppose $\mathcal{Q}$ is an ϵ-sparse quantum code. For a single round of the distillation process, if and only if the following conditions are satisfied:*

1. *The failure rate p of any noisy quantum gates is sufficiently small;*

2. *The classical error-correcting code $\mathcal{C}_{c_\mu}$ can correct $t_{c_\mu} \geq t-1$ errors;*

then the output blocks will be qualified after ideal postselection with probability at least $1 - O(\epsilon)$.

Remark 2. Here, we emphasize that the sparsity of a quantum code is technically assumed for the convenience of proof, but it may not be necessary in practice. This is supported numerically by the state distillation for the quantum Golay code in the next section. In fact, for $\mathcal{C}_{c_\mu}$ with sufficient sparse parity check matrix, (e.g., $M \leq t_{c_\mu}/2$), one can see that the probability to generate a syndrome pattern which is not good or approximately good but can pass the post-selection is less than $O(p^{t_{c_\mu}})$. In such scenario, the output ancillas are qualified. This is true for arbitrary CSS code.

Although ideal postselection is not practical, one can use a more efficient error-detecting code, as long as the code length of $\mathcal{C}_{d_1}$ ($\mathcal{C}_{d_2}$) is less than $|\mathcal{SC}_1|$ ($|\mathcal{SC}_2|$), to approximate it. Since one just needs to check parities of the estimated syndrome bits and no actual decoding is required, a broad range of classical codes with good error-detecting ability can be chosen.

In general, if we have a family of ϵ-sparse codes, ϵ will approach zero when the code length is sufficiently large. We also need to emphasize that Theorem 4 just guarantees the output states are qualified when p is small and the length of an ϵ-sparse code $\mathcal{Q}$ is sufficiently large. However, the constant behind big-O notation in Def. 1 could be large if M for the A matrix is not small enough. This fact suggests that one need to consider classical codes with sparse A matrices for FTQC purpose.

Asymptotically, the rejection rate for each round of distillation is $O(p^2)$, because at least two failures are needed to cause a rejection of the output blocks. Since the rejection rate is $O(p)$ for naive verification and other ancilla preparation schemes [30, 33], our distillation protocol is potentially more efficient in the small p regime. Suppose the rejection rates are, respectively, $R_1(p)$ and $R_2(p)$ for the two rounds of distillation. The overall yield rate for the fault-tolerant distillation protocol would be

$$\text{Yield}_{\text{FT}}(p) = \frac{k_{c_1} k_{c_2} (1 - R_{\text{FT}}(p))}{n_{c_1} n_{c_2}}, \tag{40}$$

where $R_{\text{FT}}(p) = 1 - (1 - R_1(p))(1 - R_2(p))$ is the overall block rejection rate of the fault-tolerant distillation protocol. It is possible that $R_1(p)$ and $R_2(p)$ are both negligible in the small p regime, so that one may obtain qualified ancilla states without sacrificing the overall yield rate. Compared with the yield rate of a naive verification process (Eq. (18)), our distillation protocol can achieve a yield rate *independent* of the distance of $\mathcal{Q}$, by adjusting the error-detection codes used for postselection, thus boost from $O(t^{-2})$ to $O(1)$ in practice for an $[[n, k, d = 2t + 1]]$ CSS code if appropriate family of classical codes are chosen.

V. NUMERICAL ESTIMATION: QUANTUM GOLAY CODE

In this section, we numerically study the performance of the distillation protocol proposed in Sec. IV. We will focus on the preparation of $|0\rangle_L^{\otimes k}$, since the preparation procedure for other ancilla states is essentially the same.

The first purpose of numerical simulation is to check whether the asymptotic scaling of the error weight distribution behaves correctly. That is, the probabilities of residual errors in the output blocks with weight w, $P_X(w)$ and $P_Z(w)$, should scale like $O(p^w)$ when p is sufficiently small. This will ensure the quality of the ancilla states. We also want to show that the results in the previous section hold for quantum codes of finite length.

The second purpose is to study the error performance of the distillation procedure and the overall yield rate. We may ask the effective error rate when the prepared ancillas contain only independent errors. A proper way to study this effective error rate is to approximate the error performance by some binomial distribution, where the probability parameter of that binomial distribution can be regarded as the effective error rate. Since correlated errors of higher weight are what concern us, we can quantify the distribution of the highest weight errors by a binomial distribution. Since we are preparing $|0\rangle_L^{\otimes k}$, we need to consider X errors of weight higher than t and Z errors of weight t (a Z error of weight $w > t$ is equivalent to one with $w' \leq t$ for $|0\rangle_L^{\otimes k}$). That is, for the X and Z errors, the effective error rate $p_{\text{eff}\,|\,X}$ and $p_{\text{eff}\,|\,Z}$ can be defined by:

$$P_X(w > t) = \sum_{w>t} \binom{n}{w} p_{\text{eff}\,|\,X}^{w} (1 - p_{\text{eff}\,|\,X})^{n-w}, \tag{41}$$

and

$$P_Z(w = t) = \binom{n}{t} p_{\text{eff}\,|\,Z}^{t} (1 - p_{\text{eff}\,|\,Z})^{n-t}. \tag{42}$$

Clearly, p_{eff} depends on the chosen quantum and classical codes, which makes it difficult to estimate analytically. Similarly, the yield rate is difficult to estimate, since the overall rejection rate R_{FT} depends on the gate failure rate, the structure of A_{c_1}, A_{c_2} in the parity-check matrices, and the specific error-detecting codes. However, $P_X(w > t)$ and $P_Z(w = t)$ can be directly estimated from numerical simulations.

In the following we choose our quantum code $\mathcal{Q}$ to be the $[[23, 1, 7]]$ quantum Golay code, which is constructed from the classical $[23, 12, 7]$ Golay code. Let both $\mathcal{C}_X$ and $\mathcal{C}_Z$ be the $[23, 12, 7]$ Golay code. Note that $C_X^{\perp}$ is a $[23, 11, 8]$ cyclic code, which is contained in $\mathcal{C}_X$ and $\mathcal{C}_Z$. Efficient ancilla preparation for the quantum Golay code has been extensively studied in Ref. [33] by using its permutation symmetry. But we ignore this symmetry in this paper, and focus on the performance of our protocol. The classical $[23, 12, 7]$ Golay code is the smallest *perfect* code with t larger than one, and this simplifies the study of correlated noise after distillation. This is because its decoder will always take the state back to the code space. The parity-check matrices H_X and H_Z can be represented as follows:

$$\begin{bmatrix}10000000000111110010010\\01000000000011111001001\\00100000000110001110110\\00010000000011000111011\\00001000000110010001111\\00000100000100111010101\\00000010000101101111000\\00000001000010110111100\\00000000100001011011110\\00000000010000101101111\\00000000001111100100101\end{bmatrix}. \tag{43}$$

Similarly, both logical X and Z operators of $\mathcal{Q}$ share the binary representation:

$$l = \left[00000000000101011100011\right].$$

and thus $\bar{X} = X^l$ and $\bar{Z} = Z^l$.

We need to choose classical codes $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ that can correct at least two errors to produce qualified ancillas, according to Theorem 4. Meanwhile, we hope that the number of 1s in each column of A_{c_1} and A_{c_2} is small enough so that the effect of error propagation during state distillation will not be too bad.

Here we choose two classical codes of distance five: the $[15, 7, 5]$ BCH code, and the $[5, 1, 5]$ repetition code. The parity-check matrix for the $[15, 7, 5]$ code is

$$\begin{bmatrix}100000001101000\\010000000110100\\001000000011010\\000100000001101\\000010001101110\\000001000110111\\000000101110011\\000000011010001\end{bmatrix}. \tag{44}$$

For comparison, we also study the case when $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ are codes of distance three. Two typical examples are the $[3, 1, 3]$ repetition code and the $[7, 4, 3]$ Hamming code. Codes with distance 5 will be used in Sec. V A and codes with distance 3 will be discussed in Sec. V B.

We also need two more classical codes $\mathcal{C}_{d_1}$ and $\mathcal{C}_{d_2}$ to check the compatibility of the estimated generalized syndrome bits. $\mathcal{C}_{d_1}$ and $\mathcal{C}_{d_2}$ should encode $k_{d_1} = r_Z + k = 12$ and $k_{d_2} = r_X = 11$ bits, respectively. Two very natural candidates for our purpose are the $[23, 12, 7]$ Golay code and its dual, which is a $[23, 11, 8]$ cyclic code.

Numerical simulation of a full cycle of the state distillation procedure is done as follows:

1. We first prepare $n_{c_1}(n_{c_2}+n')$ different blocks of qubits, and prepare each block in the $|0\rangle_L$ state of the quantum Golay code by a noisy encoding circuit. An optimized encoding circuit can be found in [33]. Every n_{c_1} blocks are put together to form a group. The extra n' groups of blocks are also prepared so that when some blocks are discarded in the first round, we can still keep $k_{c_1} n_{c_2}$ blocks for each group by appending extra blocks, to apply the second round distillation. The value of n' is determined by the rejection rate of the first round, and it can be optimized during simulation.

2. For $(n_{c_2} + n')$ groups, the distillation circuit of the first round is applied to each group, with noisy CNOT gates followed by noisy bitwise Z basis measurements on the first r_{c_1} blocks to get parity matrix σ_1 of the generalized syndromes of Z stabilizer generators and $\bar{Z}$.

3. Estimate the generalized syndromes $[\mathsf{s}_1]_{j,:}$ for each remaining block by classical decoding of $\mathcal{C}_{c_1}$. For each block, use the parity check matrix of the $[23, 8, 12]$ code in systematic form, H_{d_1}, to check the compatiblity of these estimated generalized syndromes. For the jth block, if $\mathsf{H}_{d_1}[\tilde{\mathsf{s}}_1]^T_{j,:} = 0$, that block is accepted; otherwise, the block will be discarded.

4. For each group of blocks, if certain blocks have been discarded, append accepted output blocks from the extra n' groups so that each group retains k_{c_1} blocks.

5. Do quantum error correction on the remaining blocks, following the recipe in Sec. III to remove X errors.

6. For each index j of the output blocks, assemble the jth block from each group to form a new group. Each new group contains n_{c_2} ancilla blocks, and there are k_{c_1} such groups in total. This step randomizes the correlated Z errors among blocks previously in the same group.

7. The second round distillation circuit is applied to each new group, with noisy CNOTs followed by noisy bitwise X basis qubit measurements on the first r_{c_2} blocks to get σ_2, the parity of generalized syndromes of the X stabilizer generators.

8. Estimate the binary vector $[\mathsf{s}_2]_{j,:}$ for each output block by classical decoding. For each output block, check if $\mathsf{H}_{d_2}[\tilde{\mathsf{s}}_2]^T_{j,:} = 0$ to decide whether the corresponding block is accepted or discarded.

9. Do quantum error correction on the remaining blocks to remove Z errors.

10. Check the weight of the remaining error and calcuate the weight distribution for both X and Z errors. Note that, since we are preparing $|0\rangle_L$ state with code distance 7, Z errors with weight larger than 3 are equivalent to Z errors with weight less than or equal to 3. So we just need to calculate the probability distribution of weight 1, 2, 3, and larger than 3 errors for X errors, and weight 1, 2, 3 for Z errors.

Note that for the last step above, one can not directly count the weights of the remaining error, because of the error degeneracy of quantum codes: two errors are equivalent if their product is a generalized stabilizer element. So certain high weight errors remaining after distillation in the simulation may actually be equivalent to low weight error. To cope with this problem, we check the generalized syndromes of the remaining error E_{rem} for each block and compare to a list of syndromes of all operators that are tensor products of X and Z with weight less than or equal to 4. If a certain operator has the same syndrome as E_{rem}, it is an equivalent error of

E_{rem} with the lowest weight, and its weight is regarded as the weight of E_{rem}. If E_{rem} does not have an equivalent error with weight less or equal to 4, its weight is higher than 4. For larger quantum codes with higher distances, the enumerating all low weight errors would be too tedious. For special states like $|0\rangle_L$, we can decode $\mathcal{C}_X^{\perp}$ after state preparation to find the minimum weight X error that shares the same syndromes with E_{rem}. The difficulty comes from the lack of an efficient decoder for $\mathcal{C}_X^{\perp}$ for a typical code $\mathcal{C}_X$. However, a decoder with worse performance may be used to approximately find the minimum weight error for low error weight. (This is just a technical difficulty in numerical benchmarking, and cause no trouble in practise, since such decoding is not necessary when real state distillation is carried out.)

A. Classical codes of distance five

1. *Combination A: [15,7,5]+[15,7,5]*

Figure 12 shows the probability of remaining X errors after distillation with weight 1, 2, 3, and larger than 3 ($P_X(w=1)$, $P_X(w=2)$, $P_X(w=3)$, and $P_X(w>3)$, respectively) using the $[15,7,5]$ classical BCH code for both rounds of distillation, with postselection. The three dashed reference lines represent the probability distributions of $O(p)$ (green), $O(p^2)$ (brown), and $O(p^3)$ (purple) are also given for comparison. It is clear that the probability of weight w X errors scales as $O(p^w)$ asymptotically for $w \leq 3$ when p is sufficiently small; when $w > 3$, since $[15,7,5]$ can only correct 2 errors, the distribution can only scale as $O(p^3)$.

Similarly, Fig. 13 shows the error weight distribution for Z errors. Note that for Z errors, we only care about those of weight less than or equal to 3. Compared with X errors, Z errors scale correctly only when $p < 0.0003$ for $w = 2, 3$. This is because after the first round of distillation, the error rate of the remaining Z errors on the output blocks is amplified, and will be beyond the error correction ability of the $[15,7,5]$ code for large values of p. Also note that the error rates for Z errors of all weights are much smaller than those of X errors of the same weight, because the remaining X errors from the first round propagate in the second round to the output blocks, and these X errors, though mostly uncorrelated, cannot be detected and corrected. Although the probability scaling up for X errors is correct, their absolute values are much higher than for Z errors. Overall, Fig. 12 and Fig. 13 qualitatively verify the validity of Theorem 4 for using the $[15,7,5]$ code for distillation, and also suggest that the $[23,8,11]$ and $[23,7,12]$ codes are sufficiently good as error detection codes $\mathcal{C}_{d_1}$ and $\mathcal{C}_{d_2}$ for the purpose of postselection.

The rejection rates in both rounds of distillation and overall yield rate are shown in Fig. 14. For all values of p considered here, the rejection rate for the X part of the distillation can be neglected, and for $p < 0.00025$, the rejection rate for the Z part is less than 5%. The overall yield rate will be around 20% in that gate failure regime, which is much more efficient than naive way verification (see Fig. 2). The rejection rates for both the X and Z parts scale as $O(p^2)$, as predicted in Theorem 1. Note that for larger distance codes like the $[[255,143,15]]$

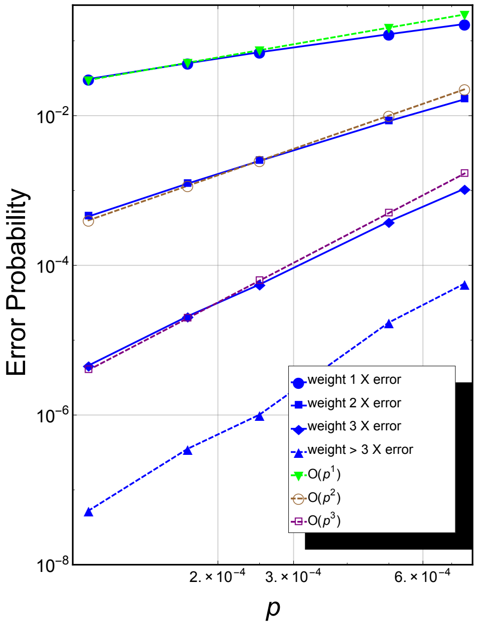

FIG. 12. (Color online) Probability of remaining X errors with weight 1, 2, 3 and larger than 3 on a single output block vs gate failure rate p, after distillation using the classical $[15,7,5]$ BCH code for both rounds, with postselection. The dashed green, brown and purple lines correspond to weight distributions $O(p)$, $O(p^2)$ and $O(p^3)$, for reference.

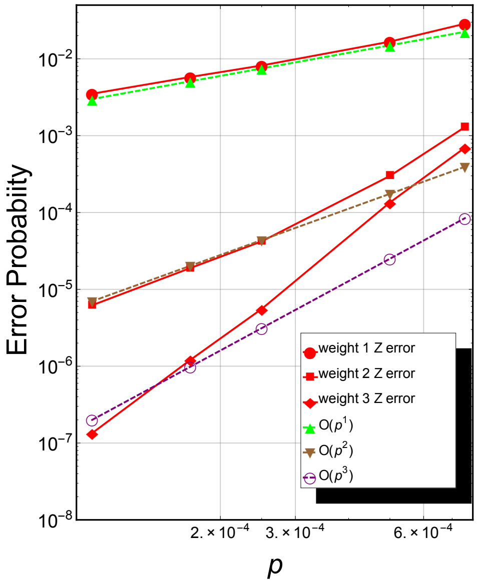

FIG. 13. (Color online) Probability of remaining of Z errors with weight 1, 2, 3 on a single output block vs gate failure rate p, after distillation using the classical $[15,7,5]$ BCH code for both rounds.

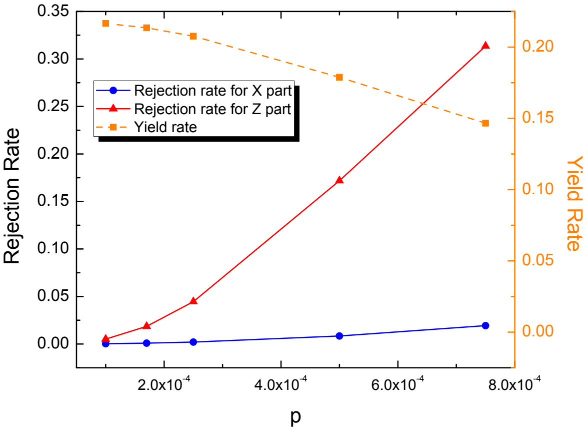

FIG. 14. (Color online) The rejection rates for both rounds of distillation, and overall yield rate, using the $[15, 7, 5]$ code in both rounds.

quantum BCH code, which is of practical interest for the large block code FTQC scheme, it would be much more efficient to use the distillation protocol than a verification circuit in the low physical failure rate region if proper $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ are chosen. This will be studied in more detail in future work.

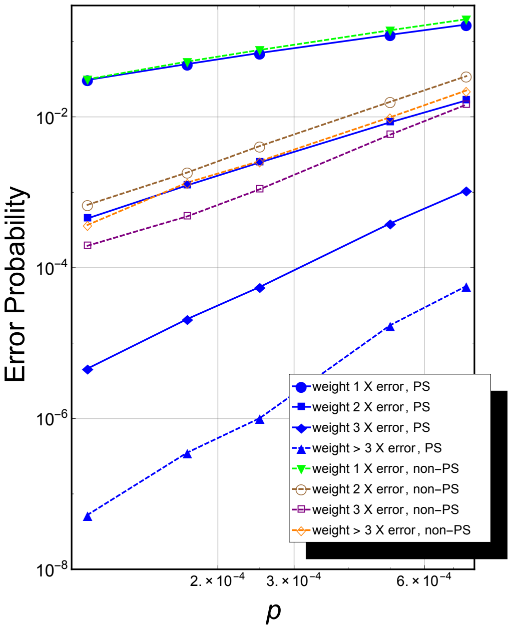

FIG. 15. (Color online) Error weight distribution after distillation for X errors, using the classical $[15, 7, 5]$ BCH code for both rounds, with and without postselection. The dashed green, brown, orange and purple lines represents the probability of remaining X errors with weights 1,2,3, and larger than 3, without postselection.

In Figs. 15 and 16 the error weight distributions with and without postselection are compared. For both parts, as stated in Theorem 1, the probability of errors of weight larger than 2 scale as $O(p^2)$ in the low failure rate region. Moreover, the probabilities of errors of weight 2 and larger than 2 remaining are close to each other for both X and Z errors, which means that the distillation protocol almost cannot reduce high weight errors at all if no postselection procedure is applied.

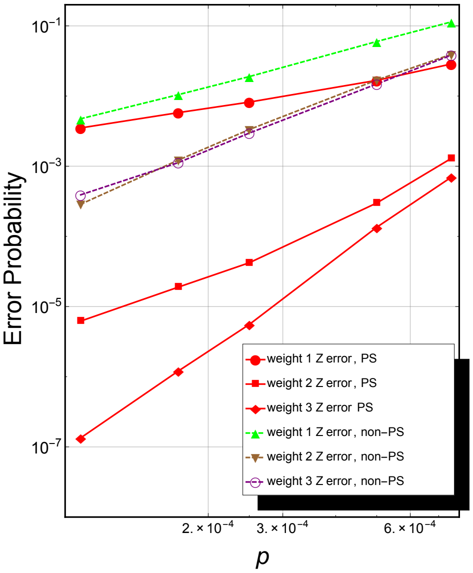

FIG. 16. (Color online) Error weight distribution after distillation for Z errors using the classical $[15, 7, 5]$ BCH code for both rounds, with and without postselection (PS). The dashed lines show the probabilities of the remaining errors without postselection.

For $p = 0.0001$, if postselection is applied, the probability of X errors with weight 3 is two orders of magnitude smaller; while for X errors with weight larger than 3, the probability is reduced by four orders of magnitude. Similarly, the probability of remaining Z errors of weight 2 is reduced by two orders of magnitude; while for weight-3 errors, the probability is reduced by three orders of magnitude. Since the rejection rate is negligibly small in this regime, postselection seems to discard almost exactly all bad ancilla blocks with only slight overkill! Thus, it can greatly improve the quality of the output ancillas with very little sacrifice of the overall yield rate.

We are also interested in the strength of the remaining errors in the low gate failure rate region, where schemes using large block codes are good enough to complete a nontrivial quantum algorithm. The effective rates for X and Z errors are $p_{\text{eff}|X} = 1.67 \times 10^{-3}$ and $p_{\text{eff}|Z} = 3.83 \times 10^{-4}$, barely good enough for the scheme in Ref. [22], where the quantum Golay code is concatenated with another code to form a large data code block. One can see that there is an asymmetry for the two types of errors: for Z errors, the effective error rate is about the same as the physical failure rate, while for X, the error rate is amplified by an order of magnitude. This seems always be the case for a two-stage distillation protocol.

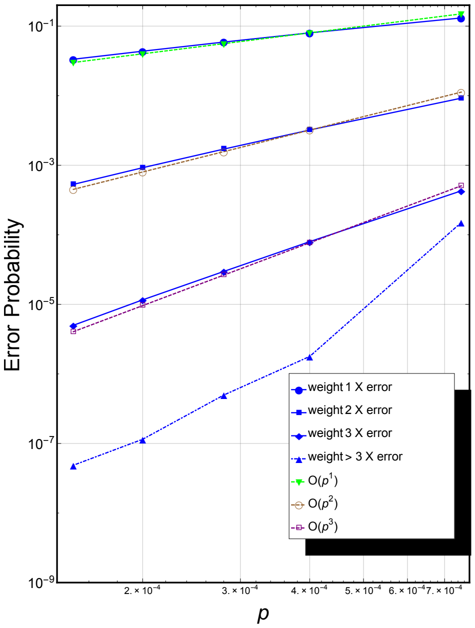

FIG. 17. (Color online) Weight distribution for X errors remaining on output blocks using the $[15, 7, 5]$ BCH code and the $[5, 1, 5]$ repetition code with postselection.

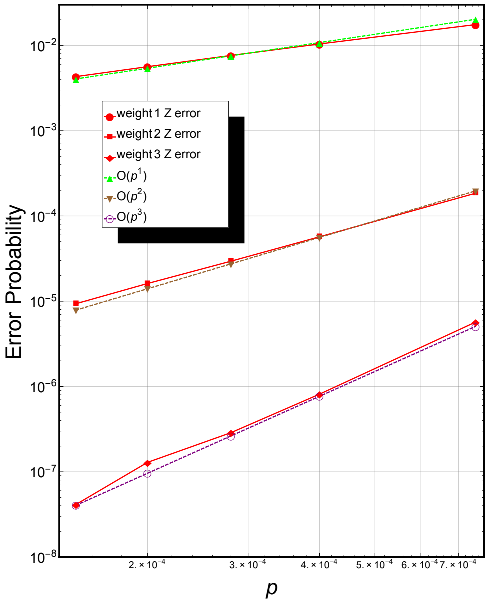

FIG. 18. (Color Online) Weight distribution of Z errors remaining on output blocks using the $[15, 7, 5]$ BCH code and the $[5, 1, 5]$ repetition code, with postselection.

2. *Combination B:[15,7,5]+[5,1,5]*

Alternatively, we can use different distance-5 classical codes $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ to see how different codes affect the quality of the output ancillas. We keep the $[15, 7, 5]$ BCH code for $\mathcal{C}_{c_1}$, and change $\mathcal{C}_{c_2}$ to be the $[5, 1, 5]$ repetition code, which has relatively larger error correction ability than the $[15, 7, 5]$ code. One might expect that the ability to suppress Z errors is even stronger in this case, which indeed turns out to be true.

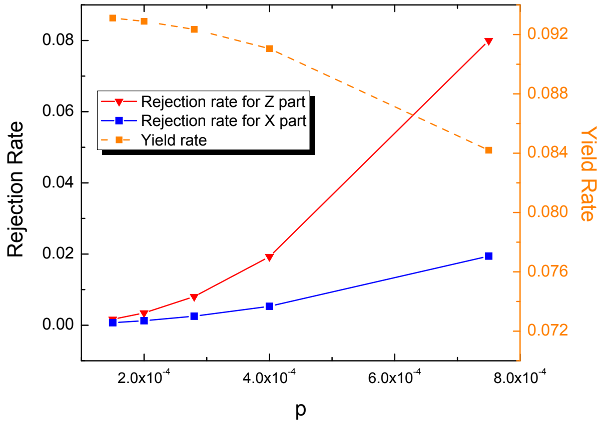

FIG. 19. (Color online) The rejection rate for both rounds of distillation and overall yield using the $[15, 7, 5]$ and $[5, 1, 5]$ codes.

Figures 17 and 18 show the weight distributions for X and Z errors with this combination of codes and Fig. 19 shows the rejection rate and overall yield rate. Compared to Combination A, the weight distribution scales correctly for all physical failure rates that we considered for both X and Z errors. The probability for a weight-2 Z error remains the same, while for weight-3 errors it can be reduced by an order of magnitude for $p < 0.00025$. However, the probabilities for residual X errors are almost the same (or slightly less) for all weight as in the previous case. This is because for the $[5, 1, 5]$ code, the corresponding A_{c_2} matrix has four 1s in a single column, which is the same for five out of the seven columns of A_{c_2} for the $[15, 7, 5]$ code (see Eq. (44): two columns contains five 1s). The rejection rate for the Z part is also reduced for all p, because only a small set of CNOT failure combinations will cause a rejection in the Z part. The overall yield rate is less than half that of Combination A, because of the low encoding rate of the $[5, 1, 5]$ code. Overall, this combination reduces the error rates for distilled blocks, but with much lower yield.

From these two examples, it seems that there is a tradeoff between the yield and the rate of effective errors remaining on the distilled blocks for small classical codes. Asymptotically, this tradeoff may disappear for large quantum codes and large classical codes with sparse parity-check matrices.

B. Classical codes of distance three

In this section, we study the case when distance-3 classical codes are used in the distillation protocol, to further verify the necessity of the conditions in Theorem 4, and complete the study of distilling $|0\rangle_L$ for the quantum Golay code.

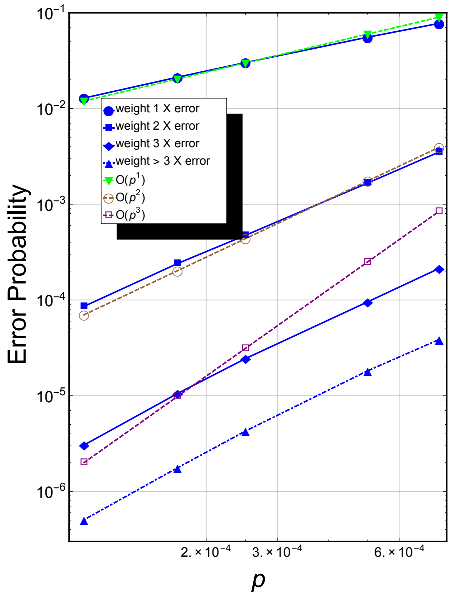

FIG. 20. (Color online) Weight distribution of X errors remaining on output blocks using the Hamming code for both rounds with postselection.

1. *Combination C: Hamming code+Hamming code*

Figures 20 and 21 show the error weight distributions for remaining X and Z errors on a single output block using the Hamming code for both rounds of distillation, with reference dashed lines representing $O(p)$, $O(p^2)$ and $O(p^3)$.

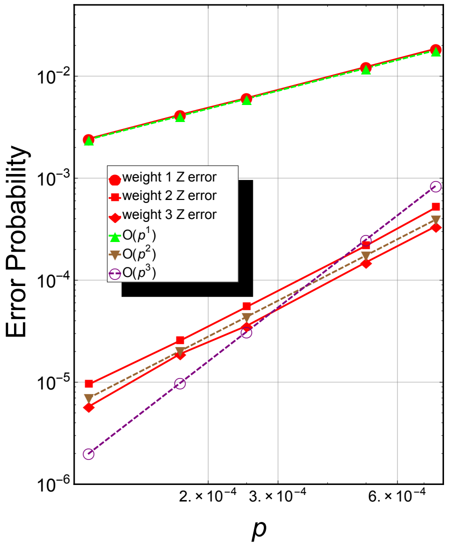

FIG. 21. (Color online) Weight distribution of Z errors remaining on output blocks using the Hamming code for both rounds with postselection.

Probabilities of both types of errors with weight larger than 2 scale as $O(p^2)$ even with postselection, as indicated by Theorem 4. The Hamming code is not able to reduce the probability of wrong syndrome estimation to $O(p^3)$. Although the postselection removes as many incompatible blocks as possible, output blocks with compatible syndromes can contain weight-3 errors with probability $O(p^2)$.

In addition, we see that the probabilities for weight 2 and 3 are close to each other for Z errors, while the probability for weight-3 X errors is much smaller than weight-2 X errors. This is because Z errors are propagated before the second stage, and the rates of syndrome flips caused by weight-2 and weight-3 Z errors are rather close.

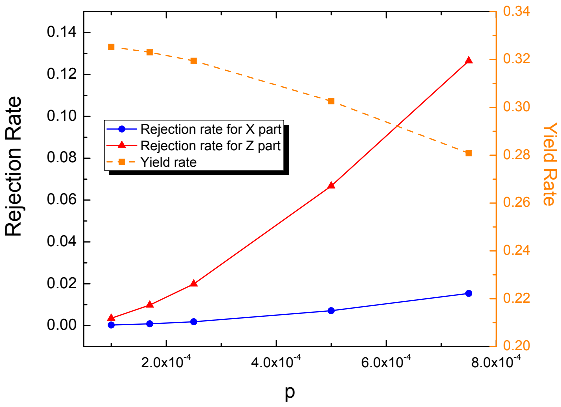

FIG. 22. (Color online) The rejection rate for both rounds of distillation and overall yield using the Hamming codes.

The low distance of the Hamming code also increases the effective error rate for both types of errors. For example, for $p = 0.0001$, $p_{\mathrm{eff}|Z} = 1.5 \times 10^{-3}$, about four times higher than for Combination A ($[15,7,5]+[15,7,5]$), while $p_{\mathrm{eff}|X} = 2.7 \times 10^{-3}$, compared with 1.67×10^{-3} for Combination A.

The rejection rate and overall yield for the Hamming code combination are shown in Fig. 22, which are comparable to those for the distance-5 code combinations. This also suggests that the $[23,8,12]$ and $[23,7,11]$ codes have removed almost all blocks with incompatible general syndromes, and postselection can make no further improvement with the limited error correction ability of classical codes. Overall, the yield is quite high, because of the high rate of the Hamming code, but the distilled states are not qualified, and the rate of the remaining effective errors is higher (especially for Z errors) than for the distance-5 code combinations studied in last subsection.

2. *Combination D: [3,1,3]+[3,1,3]*

The final classical code combination uses the $[3,1,3]$ repetition code for both rounds of distillation. It is much simpler than the Hamming code but with a lower rate. If the distillation circuit is perfect, the distilled state will have much higher fidelity compared with states distilled by Hamming codes [36].

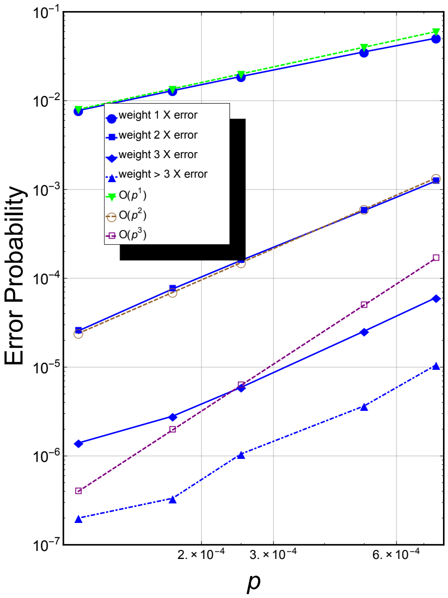

FIG. 23. (Color online) Weight distribution of X errors remaining on output blocks using the $[3,1,3]$ code for both rounds with postselection.

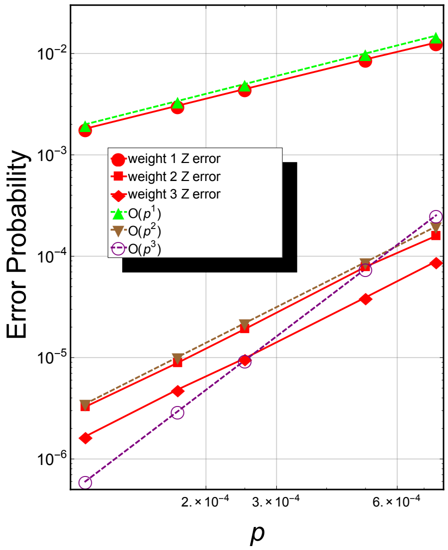

FIG. 24. (Color online) Weight distribution of Z errors remaining on output blocks using the $[3,1,3]$ code for both rounds with postselection.

Error weight distributions for both types of errors are shown in Fig. 23 and Fig. 24. As for the Hamming code, the asymptotic behavior of probability of errors with weight larger than 2 is stuck at $O(p^2)$, but the error rate is reduced because of stronger error correction ability and a simplified distillation circuit. One may also note that the difference between weight-2 and 3 Z errors is also slightly increased.

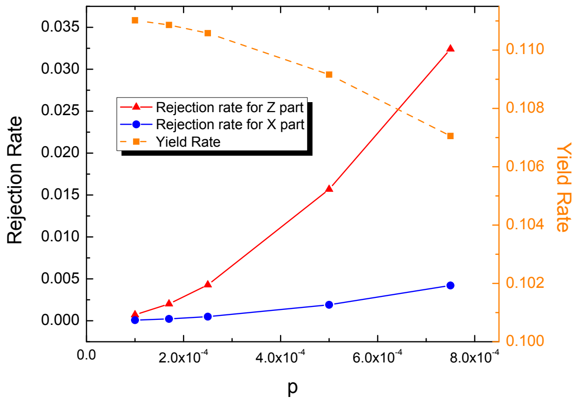

FIG. 25. (Color online) The rejection rate for both rounds of distillation and overall yield using the $[3,1,3]$ codes.

The effective error rates for X and Z errors remaining are $p_{\text{eff}\,|X} = 2.18 \times 10^{-3}$ and $p_{\text{eff}\,|Z} = 1 \times 10^{-3}$ for $p = 0.0001$, respectively, all slightly less than for the Hamming code combination. Although error propagation is not serious in this combination, the limited error correction ability of the $[3,1,3]$ code mans that $p_{\text{eff}\,|X}$ and $p_{\text{eff}\,|X}$ are still higher than for Combinations A and B. Figure 25 shows the overall yield and rejection rates for $[3,1,3] + [3,1,3]$. The rejection rates are the lowest compared with the other classical code combinations partly due to the simpler distillation circuit introducing low rates of incompatible generalized syndromes. The overall yield is also low, due to the low encoding rate.

C. Summary

We have studied four combinations of classical codes to distill $|0\rangle_L$ for the quantum Golay code. Figure 26 gives a summary, which shows the effective error rate and overall yield for these combinations for $p = 0.0001$. Two combinations of distance-5 codes (blue) sit on the left side, while two combinations of distance-3 codes (red) sit on the right side. By contrast, the distance-3 codes sit on top of the distance-5 codes. And the effective error rates for X errors are several times larger than for Z errors, because of the asymmetry of the two rounds of distillation.

A combination of classical codes to move the circle to the top-left corner is desirable for efficient FTQC. One can certainly use larger classical codes with higher rates, but such codes usually have a more complicated A matrix in the parity-check matrix, and this amplifies the effective error rate, and may increase the rejection rate to an unacceptable level, unless p is extremely small. The desired classical codes should have a sparse A matrix, strong error-correcting ability, and high rate. Certain classes of LDPC codes may possess this ability. More specifically, we would want a family of codes such that the number of 1s in each column of their A matrices

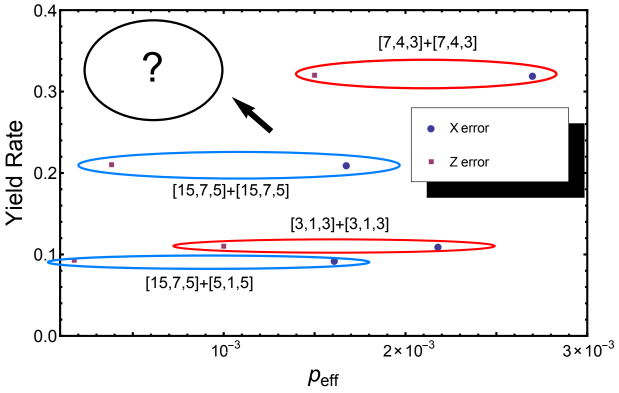

FIG. 26. (Color online) Tradeoff between effective error rate and overall yield for different combinations of classical codes for $p = 0.0001$.

is bounded by a small constant, while its error correction ability and encoding rate grow with code size. The performance of these LDPC codes is crucial for practical FTQC using large quantum block codes, and needs detailed further study. It may be possible to use the parity-check matrix H_{LDPC} of an LDPC code as the desired sparse matrix A, which will lead to a classical code with parity-check matrix $\mathsf{H} = [\mathsf{I} \mid \mathsf{H}_{\text{LDPC}}]$. This is very like the type of data-syndrome codes studied in [39], and is a subject of our future work.

VI. DISCUSSION AND FUTURE WORK

The main obstacle to implement FTQC using large block codes is the lack of an efficient and fault-tolerant preparation protocol for the required set of stabilizer ancilla states. Here, one needs not only to prepare ancillas with high fidelity, but also to suppress high-weight errors in each code block. Our new fault-tolerant distillation protocol greatly improves on the one in [36], and solves this challenge to a large extent. We use classical error-correcting codes to estimate the generalized syndrome sets, and classical error-detecting codes to pick out those ancillas with compatible estimated generalized syndromes from the output blocks, so that the quality of the output ancillas can be ensured. We proved that the output blocks are qualified if sufficiently powerful error-correcting codes and error-detecting codes are used, and the rejection rate of postselection can be very low in the low failure rate regime. Numerical simulations of fault-tolerantly preparing the $[[23, 1, 7]]$ quantum Golay code were carried out, which verify the validity of the protocol with small classical codes. The quality of the output ancillas is also greatly improved, with negligible extra cost introduced by postselection.

The error rate for residual X errors is much higher than Z errors for $|0\rangle_L$ state distillation, because of the X error propagation in the second stage of the distillation. As a result, more X errors propagate to data block than Z errors. This may suggest that we choose asymmetric CSS codes [47] as our data block codes, which can correct more X errors than Z errors, or vice versa.

Error propagation during distillation can be effectively suppressed by reducing the number of 1s in each column of the A matrices for $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$. At the same time, $\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$ should possess strong error-correcting ability and high encoding rate (so that high distillation throughput is available). For small classical codes, these requirements somewhat conflict with each other. On the other hand, medium or large-sized LDPC codes can simultaneously fulfill these requirements, and have the potential to generate qualified ancillas with low effective error rate and high throughput. This makes LDPC codes an important candidate for our protocol. The actual performance of LDPC codes may determine the overall resource overhead of large block code FTQC, and needs careful study.

For FTQC using large block codes, we are more interested in highly efficient quantum codes with large length, such as the $[[127, 57, 11]]$, or $[[255, 143, 15]]$ quantum BCH codes [48] concatenated with the quantum Golay code, or quantum LDPC codes [49, 50]. According to Theorem 4, their large t values suggest that the classical codes used in the distillation procedure should also have large distances. An extra difficulty is that these codes are usually not perfect codes, which means that the quantum decoder may not be able to take the corrupted state back to the code space if the weight of residual error is too high. So, even if the estimated generalized syndromes of the output blocks are correct, there can still exist high-weight residual errors on the output blocks after quantum error correction. This implies that we may need another layer of postselection for general quantum codes. For example, consider distillation of $|0\rangle_L^{\otimes 57}$ for the $[[127, 57, 11]]$ quantum BCH code where we just try to remove X errors. The corresponding $\mathcal{C}_X$ and $\mathcal{C}_Z$ are the $[127, 92, 11]$ classical BCH code. For the estimated generalized syndromes $\tilde{g}_Z$ obtained from classical decoding of $\mathcal{C}_{c_\mu}$, one can use the Berlekamp-Massey algorithm [51, 52] for the $[127, 92, 11]$ code to estimated the X errors, and then follow the recipe in Sec. III to remove them. Suppose the binary representation of the decoded error is $\tilde{e}$. Then, if the quantum decoding fails, in general, $\mathsf{H}_Z\tilde{e}^T \neq \tilde{g}_Z^T$. If this is true, then blocks with such estimated generalized syndromes cannot be used, and should be discarded. The cost of this extra layer of postselection needs further study.

Another potential way to further increase the efficiency of the distillation procedure is to take advantage of properties of the encoding circuits themselves [53], and the permutation symmetries of the quantum codes [33, 54]. For small quantum codes like the Steane code and Golay code, the encoding circuit may only generate a small subset of correlated errors, as observed in [53] and [33]. So it is possible to permute the qubits after encoding so that the correlated errors for differently prepared blocks are different from each other, and then verify the different prepared ancillas against each other. This has been shown to greatly reduce the resource overhead needed for ancilla verification [33]. Such properties for large block quantum codes remain unexplored, and could potentially increase the throughput of the distillation protocol without extra cost.

Our distillation protocol can simultaneously generate many copies of the *same* stabilizer ancilla states, which makes it very suitable for FT quantum simulation algorithms [55–

57]. It has been suggested in this context that a fixed small set of Clifford circuits are repeatedly applied in the algorithms. Whether these Clifford circuits as a whole can be implemented through a single logical state measurement, and whether the stabilizer ancillas needed for such measurements can be prepared by our protocol, are open problems. If the answers to these questions are positive, one may take advantage of our distillation protocol to generate a fixed, small set of high quality ancilla states fault-tolerantly, with high throughput, to implement these complicated Clifford circuits in a single step. Hopefully, this will greatly accelerate the speed of quantum simulation and reduce the resource overhead.

Finally, the distillation protocol has the potential to be generalized to prepare high fidelity codeword states of quantum error correction codes for Bosonic modes fault-tolerantly. Such codes include the GKP code [58] and binomial code [59], whose preparations are quite difficult due to the formidable complexity of the required operations [60–62].

ACKNOWLEDGMENTS

We thank Ben Reichardt, Hui-Khoon Ng and Yingkai Ouyang for useful discussions on state verification. YCZ and CYL acknowledge useful conversations with Kung-Chuan Hsu and Chung-Chin Lu on decoding algorithms for cyclic codes. This work is partially funded by the Singapore Ministry of Education (Tier-1 funding), through Yale-NUS Internal Grant IG14-LR001, and by NSF Grant No. CCF-1421078 and Grant No. MPS-1719778. TAB also acknowledges funding from an IBM Einstein Fellowship at the Institute for Advanced Study.

Appendix A: Distillation of ancilla states for FTQC

1. Ancilla States for FTQC

Preparing ancilla states that are stabilizer states is sufficient for universal FTQC. For example, these are the clean ancilla states required for FTQC in Ref. [22]:

1. $|0\rangle_L^{\otimes k}$ and $|+\rangle_L^{\otimes k}$ are used as the X and Z ancillas in Steane's syndrome extraction.

2. $|0_1,...,+_j,0_{j+1},...,0_k\rangle_L$ (or $|+_1,...,0_j,+_{j+1},...,+_k\rangle_L$). This state has all logical qubits in $|0\rangle_L$ (or $|+\rangle_L$), but the jth qubit in $|+\rangle_L$ (or $|0\rangle_L$). It is the simultaneous +1 eigenstate of logical operators $\{\bar{Z}_i, \bar{X}_j, \forall i \neq j\}$ (or $\{\bar{X}_i, \bar{Z}_j, \forall i \neq j\}$). We denote such a state as $\|Z|\bar{X}_j\rangle_L$ (or $\|X|\bar{Z}_j\rangle_L$). Here the capital Z (or X) before the vertical line means all other logical qubits but the jth one in the block are prepared in the +1 eigenstate of $\bar{Z}$ (or $\bar{X}$). If $|0\rangle_L^{\otimes k}$ and $\|X|\bar{Z}_j\rangle$ are used as the X and Z ancillas, one can measure $\bar{Z}_j$ on the data block. Similarly, if $|+\rangle_L^{\otimes k}$ and $\|Z|\bar{X}_j\rangle_L$ are used as the Z and X ancillas, one can measure $\bar{X}_j$. It is easy to check that one can simultaneously measure $\bar{X}_{i_1},\dots,\bar{X}_{i_{k'}}$ (or $\bar{Z}_{i_1},\dots,\bar{Z}_{i_{k'}}$) by using $\|Z|\bar{X}_{i_1},\dots,\bar{X}_{i_{k'}}\rangle$ (or $\|X|\bar{Z}_{i_1},\dots,\bar{Z}_{i_{k'}}\rangle$) for $k' < k$ as the X (or Z) ancilla. Similarly, one can also measure $\bar{Z}_i\bar{Z}_j$ (or $\bar{X}_i\bar{X}_j$) by preparing $\|X|\bar{Z}_i\bar{Z}_j\rangle$ or (or $\|Z|\bar{X}_i\bar{X}_j\rangle$). A logical CNOT between different logical qubits in the same data block can also be implemented in this way.

3. One also needs the entangled Bell state between the ith logical qubit in the block a of an $[[n_a,k_a]]$ code $\mathcal{Q}_a$ and the jth logical qubit in block b of an $[[n_b,k_b]]$ code $\mathcal{Q}_b$ (where $\mathcal{Q}_a$ and $\mathcal{Q}_b$ can be different CSS codes):
$$\left|\Phi_{+,ij}^{(a,b)}\right\rangle_L = \frac{1}{\sqrt{2}}\left(|0_i\rangle_L^{(a)} \otimes |0_j\rangle_L^{(b)} + |1_i\rangle_L^{(a)} \otimes |1_j\rangle_L^{(b)}\right), \quad \text{(A1)}$$
which is the simultaneous +1 eigenstate of logical $\bar{X}_i^{(a)} \otimes \bar{X}_j^{(b)}$ and $\bar{Z}_i^{(a)} \otimes \bar{Z}_j^{(b)}$. For our purpose, we set all the other logical qubits in block a and block b in $|0\rangle_L$ state, and this state is denoted as
$$\left\|Z|\Phi_{+,ij}^{(a,b)}\right\rangle_L \equiv \left\|Z^{(a)}\left|\begin{matrix}\bar{Z}_i^{(a)}\otimes\bar{Z}_j^{(b)}\\ \bar{X}_i^{(a)}\otimes\bar{X}_j^{(b)}\end{matrix}\right|Z^{(b)}\right\rangle_L . \quad \text{(A2)}$$
Similarly, one can also define
$$\left\|X|\Phi_{+,ij}^{(a,b)}\right\rangle_L \equiv \left\|X^{(a)}\left|\begin{matrix}\bar{Z}_i^{(a)}\otimes\bar{Z}_j^{(b)}\\ \bar{X}_i^{(a)}\otimes\bar{X}_j^{(b)}\end{matrix}\right|X^{(b)}\right\rangle_L \quad \text{(A3)}$$
These logical Bell states can be used to teleport logical qubits between different blocks of different quantum codes. For example, consider a joint code block consisting of both blocks a and b. One can set $\left\|Z|\Phi_{+,ij}^{(a,b)}\right\rangle_L$ as the X ancilla and $|+\rangle_L^{\otimes k_a} \otimes |+\rangle_L^{\otimes k_b}$ as the Z ancilla for the joint blocks, and then jointly measure $\bar{X}_i^{(a)} \otimes \bar{X}_j^{(b)}$. Similarly, by setting $|0\rangle_L^{\otimes k_a} \otimes |0\rangle_L^{\otimes k_b}$ as the X ancilla and $\left\|X|\Phi_{+,ij}^{(a,b)}\right\rangle_L$ as the Z ancilla, one can jointly measure $\bar{Z}_i^{(a)} \otimes \bar{Z}_j^{(b)}$. Together, one can do Bell measurements across the blocks and teleport one logical qubit of block a to the one in block b. This is particularly useful in FTQC. For instance, one can teleport a logical qubit from an efficient storage code block, but not having a fault-tolerant non-Clifford gate, to a processor code block that supports a transversal logical T gate (like the quantum Reed Muller code [63]), and then teleport it back after the logical T is finished. This technique can to get around the Eastin-Knill theorem [64, 65], and replace magic state distillation [23], and (hopefully) reduce the overall resource overhead in the computation.

4. Entangled states between the ith logical qubit of block a and jth logical qubit of block b:
$$\begin{aligned}\left|\Omega_{ij}^{(a,b)}\right\rangle_L = &\frac{1}{2}\Big(|0_i\rangle_L^{(a)} \otimes |0_j\rangle_L^{(b)} + |0_i\rangle_L^{(a)} \otimes |1_j\rangle_L^{(b)} \\ &+|1_i\rangle_L^{(a)} \otimes |0_j\rangle_L^{(b)} - |1_i\rangle_L^{(a)} \otimes |1_j\rangle_L^{(b)}\Big).\end{aligned} \quad \text{(A4)}$$
This ancilla is the joint +1 eigenstate of $\bar{X}_i^{(a)} \otimes \bar{Z}_j^{(b)}$ and $\bar{Z}_i^{(a)} \otimes \bar{X}_j^{(b)}$. One such state is
$$\left\|X^{(a)}, Z^{(b)} \mid \Omega_{ij}^{(a,b)}\right\rangle_L \equiv \left\|X^{(a)}\left|\begin{matrix}\bar{Z}_i^{(a)}\otimes\bar{X}_j^{(b)}\\ \bar{X}_i^{(a)}\otimes\bar{Z}_j^{(b)}\end{matrix}\right|Z^{(b)}\right\rangle_L , \quad \text{(A5)}$$

where the remaining logical qubits of blocks a and b are in $|+\rangle_L$ and $|0\rangle_L$, respectively. In this case, if block a is used as the Z ancilla and block b is used as the X ancilla, the Steane syndrome extraction circuit will measure $\bar{X}_i\bar{Z}_j$ on the data block, and this can be used to implement a logical Hadamard gate [22].

5. Entangled states between the jth logical qubit of block a and the jth logical qubit of block b:

$$\left|\Theta_{jj}^{(a,b)}\right\rangle_L = \frac{1}{2}\Big(|0_j\rangle_L^{(a)} \otimes |0_j\rangle_L^{(b)} + i|0_j\rangle_L^{(a)} \otimes |1_j\rangle_L^{(b)} + |1_j\rangle_L^{(a)} \otimes |0_j\rangle_L^{(b)} - i|1_j\rangle_L^{(a)} \otimes |1_j\rangle_L^{(b)}\Big). \quad \text{(A6)}$$

This state is the simultaneous +1 eigenstate of $\bar{Z}_j^{(a)} \otimes \bar{Y}_j^{(b)}$ and $\bar{X}_j^{(a)} \otimes \bar{Z}_j^{(b)}$. One such state is

$$\left\| X^{(a)}, Z^{(b)} \;\middle|\; \Theta_{jj}^{(a,b)} \right\rangle_L \equiv \left\| X^{(a)} \left| \begin{matrix} \bar{Z}_j^{(a)} \otimes \bar{Y}_j^{(b)} \\ \bar{X}_j^{(a)} \otimes \bar{Z}_j^{(b)} \end{matrix} \right| Z^{(b)} \right\rangle_L . \quad \text{(A7)}$$

This state can be used to measure logical $\bar{Y}_j$ on the data block when setting block a as the Z ancilla and block b as the X ancilla in the Steane syndrome measurement. As a byproduct of this measurement of $\bar{Y}_j$, an eigenstate of $\bar{Y}_j$ is prepared in the data block simultaneously, which can be used to implement a logical phase gate.

Note that all these stabilizer states can be produced by (imperfect) Clifford encoding circuits with CNOT gates only, together with the ability to prepare physical qubits in $|0\rangle$ and $|+\rangle$.

2. Distillation circuits

In this subsection, we generalize the distillation for $|0\rangle_L^{\otimes k}$ (or $|+\rangle_L^{\otimes k}$) so that all stabilizer states in Appendix. A 1 for FTQC can be well prepared.

$\|Z|\bar{X}_j\rangle_L$ and $\|X|\bar{Z}_j\rangle_L$— The distillation circuit in Fig. 4 can also be used to to distill $\|Z|\bar{X}_j\rangle_L$ and $\|X|\bar{Z}_j\rangle_L$. Still, there's is some difference in the decoding procedure. For $\|Z|\bar{X}_j\rangle_L$, we use $\mathcal{C}_{c_1}$ to check eigenvalues of all Z stabilizer generators and $\bar{Z}_i$ $(i \neq j)$ (together form the eigenvalues of S_1) in the first round. Eigenvalues of the X stabilizer generators and $\bar{X}_j$ (they together form the eigenvalues of S_2) are checked at the second round by $\mathcal{C}_{c_2}$ to remove Z errors. The procedure is similar for $\|X|\bar{Z}_j\rangle_L$.

$\left\|Z|\Phi_{+,ij}^{(a,b)}\right\rangle_L$ and $\left\|X|\Phi_{+,ij}^{(a,b)}\right\rangle_L$— We may also need to distill states from a pair of two quantum codes $\mathcal{Q}_a$ and $\mathcal{Q}_b$. We emphasize that $\mathcal{Q}_a$ and $\mathcal{Q}_b$ can be different. For $\left\|Z|\Phi_{+,ij}^{(a,b)}\right\rangle_L$, the distillation circuit is the same as for $|0\rangle_L$. For example, we can simultaneously remove X errors from both blocks as shown in Fig. 27. By checking the eigenvalues of the Z stabilizer generators for both blocks, and logical operators $\bar{Z}_u^{(a)}$ for $u \neq i$, $\bar{Z}_{u'}^{(b)}$ for $u' \neq j$ and $\bar{Z}_i^{(a)} \otimes \bar{Z}_j^{(b)}$, one can remove X errors on both blocks simultaneously. At the second round of distillation, eigenvalues of the X stabilizer generators and $\bar{X}_i^{(a)} \otimes \bar{X}_j^{(b)}$ are checked to remove Z errors. The process to distill $\left\|X|\Phi_{+,ij}^{(a,b)}\right\rangle_L$ for Z ancillas is similar.

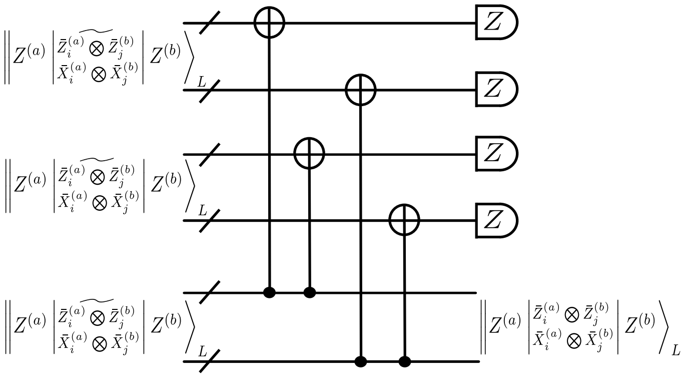
$\left\|Z^{(a)}\left|\begin{matrix}\bar{Z}_i^{(a)}\otimes\bar{Z}_j^{(b)}\\ \bar{X}_i^{(a)}\otimes\bar{X}_j^{(b)}\end{matrix}\right|Z^{(b)}\right\rangle_L$
Z
Z
Z
Z

FIG. 27. The distillation circuit for one type of the $|\Phi_{+,ij}^{(a,b)}\rangle_L$ ancilla distillation by the classical $[3,1,3]$ repetition code to remove X errors at the first round.

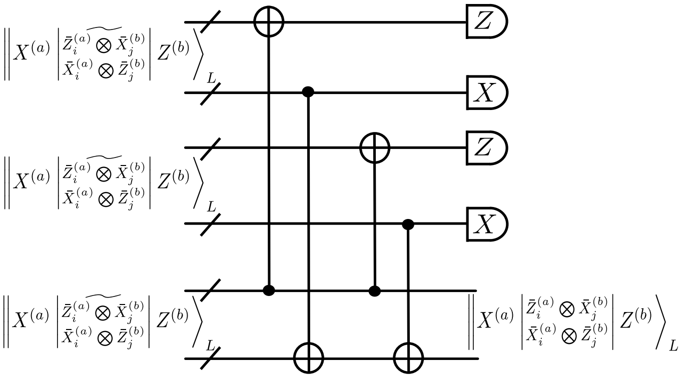
$\left\|X^{(a)}\left|\begin{matrix}\bar{Z}_i^{(a)}\otimes\bar{X}_j^{(b)}\\ \bar{X}_i^{(a)}\otimes\bar{Z}_j^{(b)}\end{matrix}\right|Z^{(b)}\right\rangle_L$
Z
X
Z
X

FIG. 28. The distillation circuit for the $|\Omega_{ij}^{(a,b)}\rangle_L$ ancilla using classical $[3,1,3]$ repetition code to remove X errors from block a and Z errors from block b.

$\left\|X^{(a)}, Z^{(b)} \;\middle|\; \Omega_{ij}^{(a,b)}\right\rangle_L$— This is another entangled state between the ith logical qubit of block a and the jth logical qubit of block b. The rest of the logical qubits for the first block are are set to $|+\rangle_L$, while for the second block, they are set to $|0\rangle_L$. Again, we need to distill the ancilla blocks in pairs. But the situation is different from the previous case, since we need to check the eigenvalues of $\bar{X}_i^{(a)} \otimes \bar{Z}_j^{(b)}$ and $\bar{Z}_i^{(a)} \otimes \bar{X}_j^{(b)}$. The state will be destroyed if both blocks are distilled to remove X or Z errors simultaneously. Instead, in the first round, we distill block a to remove X error and block b to remove Z errors, as shown in Fig. 28. Then in the second round, Z errors of block a and X errors of block b are removed. The same procedure can be used to distill more general entangled states such as

$$\left\| X^{(a)} \left| \begin{matrix} \bar{Z}_i^{(a)} \otimes \bar{X}_j^{(b)} \bar{X}_k^{(b)} \\ \bar{X}_i^{(a)} \otimes \bar{Z}_j^{(b)} \end{matrix} \right| Z^{(b)} \right\rangle_L ,$$

which is an entangled state between the ith logical qubit of block a and the jth and kth logical qubits of block b .

$\left\| X^{(a)}, Z^{(b)} \mid \Theta^{(a,b)}_{jj} \right\rangle_L$ — This entangled state between the jth logical qubit of block a and the jth logical qubit of block b is used to measure $\bar{Y}_j$ of the data block $\mathcal{Q}$ when both $\mathcal{Q}_a$ and $\mathcal{Q}_b$ are the same as $\mathcal{Q}$.

1. $\mathcal{Q}$ is a symmetric CSS codes.

2. The weights of the codewords of the underlying classical codes are $0 \mod 4$.

3. The weight of $\bar{X}_j$ is odd.

Then, after preparing the qualified ancilla state $\left\| X^{(a)}, Z^{(b)} \mid \Omega^{(a,b)}_{jj} \right\rangle_L$, we apply bitwise phase gates $P = \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{i} \end{bmatrix}$ on block b, followed by $\bar{Z}^{(b)}_j$. This procedure will give a qualified ancilla state $\left\| X^{(a)}, Z^{(b)} \mid \Theta^{(a,b)}_{jj} \right\rangle_L$. This can be shown as follows: bitwise phase gates will preserve the stabilizer group with the above properties. All other logical qubits will not be affected by the bitwise phase gates since they are in the joint +1 eigenstate of $\bar{Z}_{i'}$ $(i' \neq j)$.

Appendix B: Table of Notation

symbol	explanation
$\mathcal{C}_Z, \mathcal{C}_X$	Classical codes used to construct CSS code $\mathcal{Q}$.
$\mathcal{C}_{c_1}, \mathcal{C}_{c_2}, \mathcal{C}_{d_1}, \mathcal{C}_{d_2}$	Classical codes for syndrome estimation ($\mathcal{C}_{c_1}$ and $\mathcal{C}_{c_2}$) and compatibility ($\mathcal{C}_{d_1}$ and $\mathcal{C}_{d_2}$).
$G^{(j)}_{Z\,\vert\, i}, G^{(j)}_{X\,\vert\, i}, g^{(j)}_{Z\,\vert\, i}, g^{(j)}_{X\,\vert\, i}$	the ith Z (X) stabilizer generator of $\mathcal{Q}$ for block j and its eigenvalue in the binary form.
$L^{(j)}_i, \ell^{(j)}_i$	the ith stabilizer logical operator and its eigenvalue in the binary form.
$S^{(j)}_1, S^{(j)}_2$	Stabilizers of block j whose eigenvalues need to be estimated in the first (second) round of distillation.
$\mathcal{SC}^{(j)}_1, \mathcal{SC}^{(j)}_2$	Stabilizer group of block j generated by $S^{(j)}_1$ ($S^{(j)}_2$).
$\mathrm{SC}_{1,i}, \mathrm{SC}_{2,i}$	The ith element in $\mathcal{SC}_1$ ($\mathcal{SC}_2$) .
SE_1, SE_2	Extended stabilizers whose eigenvalues needs to be estimated for postselection.
$\mathsf{s}_\mu, \widetilde{\mathsf{s}}_\mu$ for $\mu = 1, 2$	Array of generalized syndrome bits according to S_μ, and its estimate by classical decoding.
$\mathsf{se}_\mu, \widetilde{\mathsf{se}}_\mu$ for $\mu = 1, 2$	Array of extended syndrome bits according to $S'_\mu \bigcup S_\mu$, and its estimate by classical decoding.
$\mathsf{sc}_\mu, \widetilde{\mathsf{sc}}_\mu$ for $\mu = 1, 2$	Array of all syndrome bits according to $\mathcal{SC}_\mu$, and their estimates by classical decoding of $\mathcal{C}_{c_\mu}$.
σ_μ, $\mu = 1, 2$	Parity matrices of generalized syndrome array obtained after bit-wise measurements.
t, t_c	Number of correctable errors by a quantum (classical) code.
$\mathscr{G}, \mathscr{R}$,	Failure sets that occur in the quantum circuit.
$\mathscr{G}(\tau), \mathscr{R}(\tau)$	Failures in $\mathscr{G}$ and $\mathscr{R}$ that occur at the τth time step.
$\mathscr{G}_X, \mathscr{G}_Z, \mathscr{R}_X, \mathscr{R}_Z$	Collection of the X (Z) components of the failure set $\mathscr{G}$ ($\mathscr{R}$).
$\mathscr{G}_{\mathrm{P}}, \mathscr{G}_{\mathrm{D}}$	Failures in $\mathscr{G}$ that occur in preparation and distillation stages.
$E_{\mathscr{G}}$	Effective errors caused by $\mathscr{G}$ (preceding the distillation circuit).
$\mathrm{Q}_{\mathscr{G}(\tau)}, \mathrm{Q}^{(j)}_{\mathscr{G}(\tau)}$	Support of $\mathscr{G}$ at time step τ and subset of $\mathrm{Q}_{\mathscr{G}(\tau)}$ on block j.
$\mathrm{QE}_{\mathscr{G}}, \mathrm{QE}^{(j)}_{\mathscr{G}}$	Support of $\mathscr{G}$ for CNOT failures in the distillation circuit.
$\mathrm{Q}_{E_{\mathscr{G}}}, \mathrm{Q}^{(j)}_{E_{\mathscr{G}}}$	Support of the effective errors $E_{\mathscr{G}}$.
$S^{(j)}_\mu\vert_{E_{\mathscr{G}}}, \mathcal{SC}^{(j)}_\mu\vert_{E_{\mathscr{G}}}$	Stabilizers in $S^{(j)}_\mu$ ($\mathcal{SC}^{(j)}_\mu$) that anticommute with $E_{\mathscr{G}}$.
$\mathcal{I}$	A function that extracts the indices of a set of qubits or a set of stabilizers.
$\mathcal{B}_{\mathscr{G}}$	Index set of ancilla blocks affected by $E_{\mathscr{G}}$.
$\mathcal{B}_{\mathrm{f}}$	Index set of ancilla blocks with non-zero syndromes of a good syndrome pattern.

TABLE I. Table of notation used in the paper.

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

[2] P. W. Shor, Phys. Rev. A **52**, R2493 (1995).

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

[4] F. Gaitan, *Quantum Error Correction and Fault Tolerant Quantum Computing* (CRC, 2008).

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

[6] P. Shor, in *Proc. 37th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society Press, Los Alamitos, CA, 1996) p. 56.

[7] D. Aharonov and M. Ben-Or, in *Proc. 29th Annual ACM Symposium on the Theory of Computation* (ACM Press, New York, 1997) p. 176.

[8] D. Gottesman, *Stabilizer codes and quantum error correction*, Ph.D. thesis, California Institute of Technology (1997), eprint arXiv:quant-ph/9705052.

[9] A. Kitaev, Ann. of Phys. **303**, 2 (2003).

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

[11] B. M. Terhal and G. Burkard, Phys. Rev. A **71**, 012336 (2005).

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

[13] D. Aharonov, A. Kitaev, and J. Preskill, Phys. Rev. Lett. **96**, 050504 (2006).

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

[15] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **6**, 97 (2006).

[16] A. W. Cross, D. P. Divincenzo, and B. M. Terhal, Quantum Inf. Comput. **9**, 0541 (2009).

[17] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **8**, 181 (2008).

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

[19] A. M. Steane, Phys. Rev. A **68**, 042322 (2003).

[20] A. M. Steane, Quantum Inf. Comput. **7**, 171 (2007).

[21] D. Gottesman, Quantum Inf. Comput. **14**, 1338 (2014).

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

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

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

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

[26] T. Jochym-OConnor and R. Laflamme, Phys. Rev. Lett. **112**, 010505 (2014).

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

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

[29] T. J. Yoder, R. Takagi, and I. L. Chuang, Phys. Rev. X **6**, 031039 (2016).

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

[31] B. W. Reichardt, arXiv preprint quant-ph/0406025 (2004).

[32] D. P. DiVincenzo and P. Aliferis, Phys. Rev. Lett. **98**, 020501 (2007).

[33] A. Paetznick and B. W. Reichardt, Quantum Inf. Comput. **12**, 1034 (2012).

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

[35] H. Goto, Sci. Rep. **6**, 19578 (2016).

[36] C.-Y. Lai, Y.-C. Zheng, and T. A. Brun, Phys. Rev. A **95**, 032339 (2017).

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

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

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

[40] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes* (North Holland, 1983).

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

[42] J. Benhelm, G. Kirchmair, C. F. Roos, and R. Blatt, Nature Phys. **4**, 463 (2008).

[43] Barends *et al.*, Nature. **508**, 500 (2014).

[44] Overall, it requires more than $2 + 4+, \ldots, +2t = t^2 + t$ ancilla blocks for logical 0 state verification. But for general stabilizer ancillas, it requires at least $(t+1)^2$ ancilla blocks.

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

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

[47] P. Brooks and J. Preskill, Phys. Rev. A **87**, 032310 (2013).

[48] M. Grassl and T. Beth, in *Proc. X. Int'l. Symp. Theoretical Electrical Engineering, Magdeburg* (1999) p. 207.

[49] J.-P. Tillich and G. Zemor, IEEE Trans. Inf. Theory **60**, 1193 (2014).

[50] M.-H. Hsieh, W.-T. Yen, and L.-Y. Hsu, IEEE Trans. Inf. Theory **57**, 1761 (2011).

[51] E. R. Berlekamp, *Algebraic coding theory* (McGraw-Hill New York, 1968).

[52] J. L. Massey, IEEE Trans. Inf. Theory **15**, 122 (1969).

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

[54] R. Chao and B. W. Reichardt, arXiv:1705.05365 (2017).

[55] D. Wecker, B. Bauer, B. K. Clark, M. B. Hastings, and M. Troyer, Phys. Rev. A **90**, 022305 (2014).

[56] M. B. Hastings, D. Wecker, B. Bauer, and M. Troyer, Quantum Inf. Comput. **15**, 1 (2015).

[57] D. Poulin, M. B. Hastings, D. Wecker, N. Wiebe, A. C. Doberty, and M. Troyer, Quantum Inf. Comput. **15**, 361 (2015).

[58] D. Gottesman, A. Kitaev, and J. Preskill, Phys. Rev. A **64**, 012310 (2001).

[59] M. H. Michael, M. Silveri, R. Brierley, V. V. Albert, J. Salmilehto, L. Jiang, and S. M. Girvin, Phys. Rev. X **6**, 031006 (2016).

[60] B. Travaglione and G. J. Milburn, Phys. Rev. A **66**, 052322 (2002).

[61] H. M. Vasconcelos, L. Sanz, and S. Glancy, Opt. Lett. **35**, 3261 (2010).

[62] B. Terhal and D. Weigand, Phys. Rev. A **93**, 012315 (2016).

[63] A. Steane, IEEE Trans. Inf. Theory. **45**, 1701 (1999).

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

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