Magic-state distillation with low overhead

Sergey Bravyi[1] and Jeongwan Haah[2]
[1]*IBM Watson Research Center, Yorktown Heights, New York 10598, USA*
[2]*Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA*
(Received 3 October 2012; published 27 November 2012)

We propose a family of error-detecting stabilizer codes with an encoding rate of 1/3 that permit a transversal implementation of the gate $T = \exp(-i\pi Z/8)$ on all logical qubits. These codes are used to construct protocols for distilling high-quality "magic" states $T|+\rangle$ by Clifford group gates and Pauli measurements. The distillation overhead scales as $O(\log^{\gamma}(1/\epsilon))$, where ϵ is the output accuracy and $\gamma = \log_2(3) \approx 1.6$. To construct the desired family of codes, we introduce the notion of a triorthogonal matrix, a binary matrix in which any pair and any triple of rows have even overlap. Any triorthogonal matrix gives rise to a stabilizer code with a transversal T gate on all logical qubits, possibly augmented by Clifford gates. A powerful numerical method for generating triorthogonal matrices is proposed. Our techniques lead to a twofold overhead reduction for distilling magic states with accuracy $\epsilon \sim 10^{-12}$ compared with previously known protocols.

DOI: 10.1103/PhysRevA.86.052329 PACS number(s): 03.67.Pp

I. INTRODUCTION

Quantum error-correcting codes provide a means of trading quantity for quality when unreliable components must be used to build a reliable quantum device. By combining together sufficiently many unprotected noisy qubits and exploiting their collective degrees of freedom insensitive to local errors, quantum coding allows one to simulate noiseless logical qubits and quantum gates up to any desired precision provided that the noise level is below a constant threshold value [1–4]. Protocols for fault-tolerant quantum computation with an error threshold close to 1% have been proposed recently [5–7].

An important figure of merit of fault-tolerant protocols is the cost of implementing a given logical operation such as a unitary gate or a measurement with a desired accuracy ϵ. Assuming that elementary operations on unprotected qubits have a unit cost, all fault-tolerant protocols proposed so far, including the ones based on concatenated codes [4] and topological codes [6–8], enable implementation of a universal set of logical operations with the cost $O(\log^{\beta}(1/\epsilon))$, where the scaling exponent β depends on a particular protocol.

For protocols based on stabilizer codes [9] the cost of a logical operation may also depend on whether the operation is a Clifford or a non-Clifford one. The set of Clifford operations (CO) consists of unitary Clifford group gates, such as the Hadamard gate H, the $\pi/4$ rotation $S = \exp(-i\pi Z/4)$, and the controlled-NOT (CNOT) gate, preparation of ancillary $|0\rangle$ states, and measurements in the $|0\rangle,|1\rangle$ basis. Logical CO usually have a relatively low cost as they can be implemented either transversally [9] or, in the case of topological stabilizer codes, by the code deformation method [7,8,10]. On the other hand, logical non-Clifford gates, such as the $\pi/8$ rotation $T = \exp(-i\pi Z/8)$, usually lack a transversal implementation [11,12] and have a relatively high cost that may exceed the one of CO by orders of magnitude [8]. Reducing the cost of non-Clifford gates is an important problem since the latter constitute a significant fraction of any interesting quantum circuit.

The present paper addresses this problem by constructing low overhead protocols for the magic-state distillation, a particular method of implementing logical non-Clifford gates proposed in [13]. A magic state is an ancillary resource state ψ that combines two properties.

Universality. Some non-Clifford unitary gate can be implemented using one copy of ψ and CO. The ancilla ψ can be destroyed in the process.

Distillability. An arbitrarily good approximation to ψ can be prepared by CO, given a supply of raw ancillae ρ with the initial fidelity $\langle\psi|\rho|\psi\rangle$ above some constant threshold value.

Since the Clifford group augmented by any non-Clifford gate is computationally universal [14], magic-state distillation can be used to achieve universality at the logical level provided that logical CO and logical raw ancillae ρ are readily available.

Below we shall focus on the magic state

$$|A\rangle = T|+\rangle \sim |0\rangle + e^{i\pi/4}|1\rangle .$$

A single copy of $|A\rangle$ combined with a few CO can be used to implement the T gate [15], thereby providing a computationally universal set of gates [13,16]. It was shown by Reichardt [17] that state $|A\rangle$ is distillable if and only if the initial fidelity $\langle A|\rho|A\rangle$ is above the threshold value $(1 + 1/\sqrt{2})/2 \approx 0.854$.

Our main objective will be to minimize the number of raw ancillae ρ required to distill magic states $|A\rangle$ with a desired accuracy ϵ. To be more precise, let σ be a state of k qubits which is supposed to approximate k copies of $|A\rangle$. We will say that σ has an *error rate* ϵ iff the marginal state of any qubit has an overlap of at least $1-\epsilon$ with $|A\rangle$. Suppose such a state σ can be prepared by a distillation protocol that takes as input n copies of the raw ancilla ρ and uses only CO. We will say that the protocol has a *distillation cost* $C = C(\epsilon)$ iff $n \leqslant Ck$. For example, the original distillation protocol of Ref. [13] based on the 15-qubit Reed-Muller code has a distillation cost $O(\log^{\gamma}(1/\epsilon))$, where $\gamma = \log_3(15) \approx 2.47$.

II. SUMMARY OF RESULTS

Our main result is a family of distillation protocols for state $|A\rangle$ with a distillation cost $O(\log^{\gamma}(1/\epsilon))$, where $\gamma = \log_2(\frac{3k+8}{k})$ and k is an arbitrary even integer. By choosing large enough k the scaling exponent γ can be made arbitrarily close to $\log_2(3) \approx 1.6$. The protocol works by concatenating an elementary subroutine that takes as input $3k+8$ magic states with an error rate p and outputs k magic states with an error rate $O(p^2)$. For comparison, the protocol found by Meier

et al. [18] has a distillation cost as above with the scaling exponent $\gamma = \log_2(5) \approx 2.32$. Distillation protocols with the scaling exponent $\gamma = 2$ were recently discovered by Campbell *et al.* [19], who studied extensions of stabilizer codes, CO, and magic states to qudits. We conjecture that the scaling exponent γ cannot be smaller than 1 for *any* distillation protocol and give some arguments in support of this conjecture in Sec. VI.

Our distillation scheme borrows two essential ideas from Refs. [13,18]. First, as proposed in [13], we employ stabilizer codes that admit a special symmetry in favor of transversal T gates and measure the syndrome of such codes to detect errors in the input magic states. Secondly, as proposed by Meier *et al.* [18], we reduce the distillation cost significantly by using distance-2 codes with multiple logical qubits. The new ingredient is a systematic method of constructing stabilizer codes with the desired properties. To this end we introduce the notion of a triorthogonal matrix, a binary matrix in which any pair and any triple of rows have even overlap. We show that any triorthogonal matrix G with k odd-weight rows can be mapped to a stabilizer code with k logical qubits that admit a transversal T gate on all logical qubits, possibly augmented by Clifford gates. Each even-weight row of G gives rise to a stabilizer which is used in the distillation protocol to detect errors in the input magic states. Finally, we propose a powerful numerical method for generating triorthogonal matrices. To illustrate its usefulness, we construct the first example of a distance-5 code with a transversal T gate that encodes 1 qubit into 49 qubits.

While the asymptotic scaling of the distillation cost is of great theoretical interest, its precise value in the nonasymptotic regime may offer valuable insights on the practicality of a given protocol. Using raw ancillae with the initial error rate 10^{-2} and the target error rate ϵ between 10^{-3} and 10^{-30}, we computed the distillation cost $C(\epsilon)$ numerically for the optimal sequence composed of the 15-to-1 protocol of Ref. [13], and the 10-to-2 protocol of Ref. [18]. Combining these protocols with the ones discovered in the present paper, we observed a twofold reduction of the distillation cost for $\epsilon = 10^{-12}$ and a noticeable cost reduction for the entire range of ϵ (see Table I in Sec. VIII).

Since a magic-state distillation is meant to be performed at the logical level of some stabilizer code, throughout this paper we assume that CO themselves are perfect. Whether or not this simplification is justified depends on the chosen code. More precisely, let the cost of implementing logical CO and the distillation cost be $\log^{\beta}(1/\epsilon)$ and $\log^{\gamma}(1/\epsilon)$, respectively, where ϵ is the desired precision. In the case $\beta < \gamma$, high-quality CO are cheap, and one can safely assume that CO are perfect. The opposite case, when high-quality CO are expensive (i.e., $\beta > \gamma$), is realized, for example, in the topological one-way quantum computer based on the three-dimensional cluster state introduced by Raussendorf *et al.* [8], where $\beta = 3$. As was pointed out in [8], in this case it is advantageous to use expensive high-quality CO only at the final rounds of distillation and to use relatively cheap noisy CO for the initial rounds. Using the 15-to-1 distillation protocol of Ref. [13] with $\gamma = \log_3 15 \approx 2.47$, the authors of Ref. [8] showed how to implement a universal set of logical gates with the cost $O(\log^3(1/\epsilon))$. A detailed analysis of errors in logical CO was performed by Jochym-O'Connor *et al.* [20].

The rest of the paper is organized as follows. We begin with the definition of triorthogonal matrices and state their basic properties in Sec. III. The correspondence between triorthogonal matrices and stabilizer codes with a transversal T gate is described in Sec. IV. We introduce our distillation protocols for magic state $|A\rangle$ in Secs. V and VI and Appendix A. A family of distance-2 codes with an encoding rate of $1/3$ that admit a transversal T gate is presented in Sec. VII. We compute the distillation cost of the new protocols and make a comparison with the previously known protocols in Sec. VIII. A numerical method of generating triorthogonal matrices is presented in Sec. IX. Finally, Appendix B presents the [[49,1,5]] code with a transversal T gate.

Notation. Below we adopt standard notation and terminology pertaining to quantum stabilizer codes [21]. Given a pair of binary vectors $f,g \in \mathbb{F}_2^n$, let $(f,g) = \sum_{j=1}^n f_j g_j \pmod 2$ be their inner product and $|f|$ be the weight of f, that is, the number of nonzero entries in f. Given a linear space $\mathcal{G} \subseteq \mathbb{F}_2^n$, its dual space $\mathcal{G}^{\perp}$ consists of all vectors $f \in \mathbb{F}_2^n$ such that $(f,g) = 0$ for any $g \in \mathcal{G}$. We shall use the notation X,Y,Z for the single-qubit Pauli operators. Given any single-qubit operator O and a binary vector $f \in \mathbb{F}_2^n$, the tensor product $O^{f_1} \otimes \cdots \otimes O^{f_n}$ will be denoted $O(f)$. In particular, $X(f)Z(g) = (-1)^{(f,g)} Z(g)X(f)$. The Pauli group $\mathcal{P}_n$ consists of n-qubit Pauli operators $i^{\omega} P_1 \otimes \cdots \otimes P_n$, where $P_j \in \{I,X,Y,Z\}$, and $\omega \in \mathbb{Z}_4$. The Clifford group $\mathcal{C}_n$ consists of all unitary operators U such that $U\mathcal{P}_n U^{\dagger} = \mathcal{P}_n$. It is well known that $\mathcal{C}_n$ is generated by one-qubit gates $H = (X+Z)/\sqrt{2}$ (the Hadamard gate), $S = \exp(i\pi Z/4)$ (the S gate), and the controlled-Z gate $\Lambda(Z) = \exp(i\pi\,|11\rangle\langle 11|)$. All quantum codes discussed in this paper are of the Calderbank-Shor-Steane (CSS) type [22,23]. Given a pair of linear spaces $\mathcal{F},\mathcal{G} \subset \mathbb{F}_2^n$ such that $\mathcal{F} \subseteq \mathcal{G}^{\perp}$, the corresponding CSS code has stabilizer group $\{X(f)Z(g), \quad f \in \mathcal{F},\ g \in \mathcal{G}\}$ and will be denoted as $\mathrm{CSS}(X,\mathcal{F};Z,\mathcal{G})$.

III. TRIORTHOGONAL MATRICES

To describe our distillation protocols let us define a new class of binary matrices.

Definition 1. A binary matrix G of size $m \times n$ is called triorthogonal iff the supports of any pair and any triple of its rows have even overlap, that is,

$$\sum_{j=1}^{n} G_{a,j} G_{b,j} = 0 \pmod 2 \tag{1}$$

for all pairs of rows $1 \leqslant a < b \leqslant m$ and

$$\sum_{j=1}^{n} G_{a,j} G_{b,j} G_{c,j} = 0 \pmod 2 \tag{2}$$

for all triples of rows $1 \leqslant a < b < c \leqslant m$.

An example of a triorthogonal matrix of size 5×14 is

$$G = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & & & & & & & \\ & & & & & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & & 1 & & 1 & & 1 & 1 & & 1 & & 1 & & 1 \\ & 1 & 1 & & & 1 & 1 & & 1 & 1 & & & 1 & 1 \\ & & & 1 & 1 & 1 & 1 & & & & 1 & 1 & 1 & 1 \end{bmatrix}, \tag{3}$$

where only nonzero matrix elements are shown. The two submatrices of G formed by even-weight and odd-weight rows will be denoted G_0 and G_1 respectively. The submatrix G_0 is highlighted in bold in Eq. (3). We shall always assume that G_1 consists of the first k rows of G for some $k \geqslant 0$. Define linear subspaces $\mathcal{G}_0, \mathcal{G}_1, \mathcal{G} \subseteq \mathbb{F}_2^n$ spanned by the rows of G_0, G_1, and G, respectively. Using Eq. (1) alone, one can easily prove the following.

Lemma 1. Suppose G is triorthogonal. Then (i) all rows of G_1 are linearly independent over $\mathbb{F}_2$, (ii) $\mathcal{G}_0 \cap \mathcal{G}_1 = 0$, (iii) $\mathcal{G}_0 = \mathcal{G} \cap \mathcal{G}^\perp$, and (iv) $\mathcal{G}_0^\perp = \mathcal{G}_1 \oplus \mathcal{G}^\perp$.

Proof. Let $f^1, \ldots, f^m$ be the rows of G such that the first k rows form G_1. By definition, any vector $f \in \mathcal{G}_1$ can be written as $f = \sum_{a=1}^k x_a f^a$ for some $x_a \in \mathbb{F}_2$. From Eq. (1) we infer that $(f^a, f^b) = \delta_{a,b}$ for all $1 \leqslant a,b \leqslant k$ and $(f^a, g) = 0$ for any $g \in \mathcal{G}_0$. Hence $x_a = (f, f^a)$. If $f = 0$ or $f \in \mathcal{G}_0$, then $x_a = 0$ for all a. This proves (i) and (ii). Since any row of G_0 is orthogonal to itself and any other row of G, we get $(f,g) = 0$ for all $f \in \mathcal{G}_0$ and $g \in \mathcal{G}$. This implies $\mathcal{G}_0 \subseteq \mathcal{G} \cap \mathcal{G}^\perp$. If $f = \sum_{a=1}^m x_a f^a \in \mathcal{G} \cap \mathcal{G}^\perp$, then $x_a = (f, f^a) = 0$ for all $1 \leqslant a \leqslant k$, that is, $f \in \mathcal{G}_0$. This proves (iii). Finally, (iv) follows from $\mathcal{G}_1 \oplus \mathcal{G}^\perp \subseteq \mathcal{G}_0^\perp$, $\mathcal{G}_1 \cap \mathcal{G}^\perp = 0$, and dimension counting. ■

As we show in Sec. IV, any binary matrix G with n columns and k odd-weight rows satisfying Eq. (1) gives rise to a stabilizer code encoding k qubits into n qubits. Condition (2) ensures that this code has the desirable transversality properties, namely, the encoded $|A^{\otimes k}\rangle$ state can be prepared by applying the transversal T gate $T^{\otimes n}$ to the encoded $|+^{\otimes k}\rangle$, possibly augmented by some Clifford operator. To state this more formally, define n-qubit unnormalized states

$$|G_0\rangle = \sum_{g \in \mathcal{G}_0} |g\rangle\,, \quad |G\rangle = \sum_{g \in \mathcal{G}} |g\rangle\,. \tag{4}$$

Define also a state

$$|\overline{A^{\otimes k}}\rangle = \prod_{a=1}^k [I + e^{i\pi/4} X(f^a)]\; |G_0\rangle\,, \tag{5}$$

where $f^1, \ldots, f^k$ are the rows of G_1.

Lemma 2. Suppose a matrix G is triorthogonal. Then there exists a Clifford group operator U composed of $\Lambda(Z)$ and S gates only such that

$$|\overline{A^{\otimes k}}\rangle = U T^{\otimes n} |G\rangle\,. \tag{6}$$

Proof. Below we promote the elements of binary field $\mathbb{F}_2$ to the normal integers of $\mathbb{Z}$; we associate $\mathbb{F}_2 \ni 0 \mapsto 0 \in \mathbb{Z}$ and $\mathbb{F}_2 \ni 1 \mapsto 1 \in \mathbb{Z}$. Unless otherwise denoted by (mod2) or (mod4), every sum is the usual sum for integers, and no modulo reduction is performed.

When $y = (y_1, \ldots, y_m)$ is a string of 0 or 1, let $\epsilon(y) \equiv |y| \pmod 2$ be the parity of y. Let us derive a formula for a phase factor $e^{i\pi\epsilon(y)/4}$ as a function of components y_a. Observe that

$$\epsilon(y) = \frac{1}{2}[1 - (1-2)^{|y|}] = \sum_{p=1}^{|y|} \binom{|y|}{p} (-2)^{p-1}. \tag{7}$$

Since the binomial coefficient $\binom{|y|}{p}$ is the number of ways to choose p nonzero components of y, we may write

$$e^{i\pi\epsilon(y)/4} = \exp\left[\frac{i\pi}{4}\sum_{a=1}^m y_a - \frac{i\pi}{2}\sum_{a<b} y_a y_b + i\pi \sum_{a<b<c} y_a y_b y_c\right]. \tag{8}$$

By definition of state $|G\rangle$, one has

$$T^{\otimes n}\,|G\rangle = \sum_{f \in \mathcal{G}} e^{i\pi|f|/4}\; |f\rangle\,.$$

Since $|G\rangle$ depends on the linear space $\mathcal{G}$ rather than the matrix presentation G, we may assume that all rows of G are linearly independent over $\mathbb{F}_2$. Let $g^1, \ldots, g^m$ be the rows of G, and decompose $f = \sum_{a=1}^m x_a g^a \pmod 2$, where $x_a \in \{0,1\}$ are uniquely determined by f.

Each component f_j of f is the parity of the bit string $(x_1 g_j^1, x_2 g_j^2, \ldots, x_m g_j^m)$, and $|f|$ is the sum of f_j. Hence, Eq. (8) implies

$$e^{i\pi|f|/4} = \exp\left[\frac{i\pi}{4}\sum_{a=1}^m x_a |g^a| - \frac{i\pi}{2}\sum_{a<b} x_a x_b |g^a \cdot g^b| + i\pi \sum_{a<b<c} x_a x_b x_c |g^a \cdot g^b \cdot g^c|\right], \tag{9}$$

where $g^a \cdot g^b$ denotes the bitwise AND operation. Triorthogonality condition (2) implies that the triple overlap $|g^a \cdot g^b \cdot g^c|$ is even, so we may drop the last term in Eq. (9). This is, in fact, one of the main reasons why we consider triorthogonal matrices.

Let the first k rows of G have odd weight and all others have even weight, and put

$$|g^a| = \begin{cases} 2\Gamma_a + 1 & \text{if} \quad 1 \leqslant a \leqslant k, \\ 2\Gamma_a & \text{otherwise.} \end{cases}$$

In addition, Eq. (1) implies for distinct a,b that

$$|g^a \cdot g^b| = 2\Gamma_{ab}.$$

Here all Γ_a and Γ_{ab} are integers. Thus

$$e^{i\pi|f|/4} = \exp\left[\frac{i\pi}{4}\sum_{a=1}^k x_a\right] \exp\left[\frac{i\pi}{2} Q(x_1, \ldots, x_m)\right],$$

where

$$Q(x) = \sum_{a=1}^m \Gamma_a\, x_a - 2\sum_{a<b} \Gamma_{ab}\, x_a x_b.$$

Let us show that the unwanted phase factor $e^{i\pi Q/2}$ can be canceled by a unitary Clifford operator that uses only $\Lambda(Z)$ and S gates. To this end, we rewrite $Q(x)$ as a function of f. As noted earlier, x_a are uniquely determined by f. Indeed,

there is a matrix B over $\mathbb{F}_2$ such that $x_a = \sum_p B_{ap} f_p \pmod 2$ since $\{g^a\}$ is a basis of the linear space $\mathcal{G}$. (There could be many such B.) We again use Eq. (7) with the observation that x_a is the parity of the bit string $(B_{a1} f_1, \ldots, B_{an} f_n)$ to infer

$$x_a = \sum_p B_{ap} f_p - 2 \sum_{p<q} B_{ap} B_{aq} f_p f_q \pmod 4,$$
$$2 x_a x_b = 2 \sum_{p,q} B_{ap} B_{bq} f_p f_q \pmod 4$$

for all $a,b = 1, \ldots, m$. Therefore, we can express $Q(x)$ as

$$Q(x(f)) = \sum_{p=1}^{n} \Lambda_p f_p - 2 \sum_{p<q} \Lambda_{pq} f_p f_q \pmod 4,$$

where Λ_p, Λ_{pq} are some integers determined by B, Γ_a, and Γ_{ab}, all of which depend only on our choice of the matrix G. Explicitly, $\Lambda_p = \sum_a \Gamma_a B_{ap} - 2 \sum_{a<b} \Gamma_{ab} B_{ap} B_{bp}$ and $\Lambda_{pq} = \sum_a \Gamma_a B_{ap} B_{aq} - \sum_{a<b} \Gamma_{ab} (B_{ap} B_{bq} + B_{bp} B_{aq})$.

The extra phase factor $e^{i\pi Q/2}$ is canceled by applying the $\Lambda(Z)^{\Lambda_{pq}}$ gate for each pair of qubits $p < q$ and the gate $(S^\dagger)^{\Lambda_p}$ to every qubit p. This defines the desired Clifford operator U composed of $\Lambda(Z)$ and S gates such that

$$U T^{\otimes n} |f\rangle = \exp\left[\frac{i\pi}{4} \sum_{a=1}^{k} x_a\right] |f\rangle \tag{10}$$

for all $f = \sum_{a=1}^{m} x_a g^a \pmod 2 \in \mathcal{G}$. Therefore,

$$U T^{\otimes n} |G\rangle = \prod_{a=1}^{k} [I + e^{i\pi/4} X(g^a)] |G_0\rangle = |\overline{A^{\otimes k}}\rangle. \qquad \blacksquare$$

For the later use let us state the following simple fact.

Lemma 3. Let G be a triorthogonal matrix without zero columns. If G_1 is not empty and G_0 has fewer than three rows, then G_0 must have at least one zero column.

Proof. Suppose, on the contrary, all columns of G_0 are nonzero. If G_0 has only one row, it must be the all-ones vector 1^n. Then, the inner product between 1^n and any row f of G_1 is the weight of f modulo 2, which is odd. But, the orthogonality equation (1) requires it to be even. This is a contradiction.

Suppose now that G_0 has two rows g_1, g_2. By permuting the columns we may assume that $G_0 = \begin{bmatrix} A & B & C \end{bmatrix}$, where

$$A = \begin{bmatrix} 1 \cdots 1 \\ 0 \cdots 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \cdots 0 \\ 1 \cdots 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \cdots 1 \\ 1 \cdots 1 \end{bmatrix}.$$

Choose an odd-weight row f of G_1, and let w_A, w_B, w_C be the weight of f restricted to the columns of A, B, C, respectively. The (tri)orthogonality equations (1) and (2) imply

$$|g_1 \cdot f| = w_A + w_C = 0 \pmod 2,$$
$$|g_2 \cdot f| = w_B + w_C = 0 \pmod 2,$$
$$|g_1 \cdot g_2 \cdot f| = w_C = 0 \pmod 2.$$

This is a contradiction since $|f| = w_A + w_B + w_C = 1 \pmod 2$. $\blacksquare$

IV. STABILIZER CODES BASED ON TRIORTHOGONAL MATRICES

Given a triorthogonal matrix G with k odd-weight rows, define a stabilizer code CSS $(X, \mathcal{G}_0; Z, \mathcal{G}^\perp)$ with X-type stabilizers $X(f)$, $f \in \mathcal{G}_0$, and Z-type stabilizers $Z(g)$, $g \in \mathcal{G}^\perp$. The inclusion $\mathcal{G}_0 \subseteq \mathcal{G}$ implies that all stabilizers pairwise commute.

Lemma 4. The code CSS $(X, \mathcal{G}_0; Z, \mathcal{G}^\perp)$ has k logical qubits. Its logical Pauli operators can be chosen as

$$\overline{X}_a = X(f^a), \quad \overline{Z}_a = Z(f^a), \quad a = 1, \ldots, k, \tag{11}$$

where $f^1, \ldots, f^k$ are the rows of G_1. The states $|G_0\rangle$, $|G\rangle$, and $|\overline{A^{\otimes k}}\rangle$ defined in Eqs. (4) and (5) coincide with encoded states $|0^{\otimes k}\rangle$, $|+^{\otimes k}\rangle$, and $|A^{\otimes k}\rangle$, respectively.

Proof. Indeed, the assumption that f^a have odd weight and Eq. (1) ensure that the operators defined in Eq. (11) obey the correct commutation rules, that is, $\overline{X}_a \overline{Z}_b = (-1)^{\delta_{a,b}} \overline{Z}_b \overline{X}_a$. It remains to be checked that $\overline{X}_a$ and $\overline{Z}_a$ commute with all stabilizers. Given any Z-type stabilizer $Z(g)$, $g \in \mathcal{G}^\perp$, one has $X(f^a) Z(g) = (-1)^{(f^a, g)} Z(g) X(f^a) = Z(g) X(f^a)$ since $f^a \in \mathcal{G}$ and $g \in \mathcal{G}^\perp$. Given any X-type stabilizer $X(f)$, $f \in \mathcal{G}_0$, one has $Z(f^a) X(f) = (-1)^{(f^a, f)} X(f) Z(f^a) = X(f) Z(f^a)$ since $f^a \in \mathcal{G}$ and $\mathcal{G}_0 \subseteq \mathcal{G}^\perp$; see Lemma 1. This shows that $\overline{X}_a$ and $\overline{Z}_a$ are indeed logical Pauli operators on k encoded qubits.

Property (iii) of Lemma 1 implies that $Z(g) |f\rangle = |f\rangle$ for any $f \in \mathcal{G}_0$ and any $g \in \mathcal{G} + \mathcal{G}^\perp$. Thus the state $|G_0\rangle$ defined in Eq. (4) coincides with the encoded $|0^{\otimes k}\rangle$ state. It follows that $|G\rangle = \prod_{a=1}^{k} (I + \overline{X}_a) |G_0\rangle$ is the encoded $|+^{\otimes k}\rangle$ state, while $|\overline{A^{\otimes k}}\rangle = \prod_{a=1}^{k} (I + e^{i\pi/4} \overline{X}_a) |G_0\rangle$ is the encoded $|A^{\otimes k}\rangle$ (ignoring the normalization). $\blacksquare$

Using Lemma 4 one can show that the operator $U T^{\otimes n}$ defined in Lemma 2 implements an encoded T gate on each logical qubit of the code CSS$(X, \mathcal{G}_0; Z, \mathcal{G}^\perp)$. Indeed, for any $x \in \mathbb{F}_2^k$, the encoded state $|x\rangle \equiv |x_1, \ldots, x_k\rangle$ is

$$|\overline{x}\rangle = \overline{X}_1^{x_1} \cdots \overline{X}_k^{x_k} |G_0\rangle = \sum_{f \in \mathcal{G}_0 + x_1 f^1 + \cdots + x_k f^k} |f\rangle.$$

Using Eq. (10) from the proof of Lemma 2, one arrives at

$$U T^{\otimes n} |\overline{x}\rangle = e^{i \frac{\pi}{4} \sum_{a=1}^{k} x_a} |\overline{x}\rangle.$$

This provides a generalization of a transversal T gate to multiple logical qubits.

V. DISTILLATION SUBROUTINE

We are now ready to describe the elementary distillation subroutine. It takes as input n copies of a (mixed) one-qubit ancilla ρ such that $\langle A| \rho |A\rangle = 1 - p$. We shall refer to p as the *input error rate*. Define single-qubit basis states $|A_0\rangle \equiv |A\rangle$ and $|A_1\rangle \equiv Z |A\rangle$. We shall assume that ρ is diagonal in the A basis; that is,

$$\rho = (1 - p) |A_0\rangle \langle A_0| + p |A_1\rangle \langle A_1|. \tag{12}$$

This can always be achieved by applying operators I and $A \equiv e^{-i\pi/4} SX$ with a probability of $1/2$ each to every copy

of ρ. Note that $A\,|A_\alpha\rangle = (-1)^\alpha\,|A_\alpha\rangle$; that is, the random application of A is equivalent to the dephasing in the A basis, which destroys the off-diagonal matrix elements $\langle A_0|\,\rho\,|A_1\rangle$ without changing the fidelity $\langle A_0|\,\rho\,|A_0\rangle$.

Define linear maps

$$\mathcal{T}(\eta) = T\eta T^\dagger, \quad \mathcal{E}(\eta) = (1-p)\eta + pZ\eta Z \tag{13}$$

describing the ideal T gate and the Z error, respectively. Using Clifford operations and one copy of ρ as in Eq. (12), one can implement a noisy version of the T gate, namely, $\mathcal{E}\circ\mathcal{T}$. A circuit implementing $\mathcal{E}\circ\mathcal{T}$ is shown in Fig. 1, where the Z error $\mathcal{E}$ is shown by the Z-gate box with a subscript p indicating the error probability. One can easily show that this circuit indeed implements $\mathcal{E}\circ\mathcal{T}$ by commuting $\mathcal{E}$ through the CNOT gate and the classically controlled SX gate.

The entire subroutine is illustrated in Fig. 2. The first step is to prepare k copies of the state $|+\rangle$ and encode them using the code CSS $\left(X,\mathcal{G}_0; Z,\mathcal{G}^\perp\right)$. This results in the state $|G\rangle$ defined in Eq. (4) and requires only CO.

State $|G\rangle$ is then acted upon by the map $(\mathcal{E}\circ\mathcal{T})^{\otimes n}$. The latter can be implemented using CO and n copies of ρ, as shown on Fig. 1. This results in a state

$$\eta_1 \equiv (\mathcal{E}\circ\mathcal{T})^{\otimes n}(|G\rangle\langle G|) = \mathcal{E}^{\otimes n}(\hat{T}|G\rangle\langle G|\hat{T}^\dagger),$$

where $\hat{T}\equiv T^{\otimes n}$. Next, we apply the Clifford unitary operator U constructed in Lemma 2. Since U involves only $\Lambda(Z)$ and S gates, it commutes with any Z-type error. Hence the state prepared at this point is

$$\eta_2 \equiv U\eta U^\dagger = \mathcal{E}^{\otimes n}(U\hat{T}|G\rangle\langle G|\hat{T}^\dagger U^\dagger) = \mathcal{E}^{\otimes n}(|\overline{A^{\otimes k}}\rangle\langle\overline{A^{\otimes k}}|),$$

where we have used Eq. (6). The next step is a nondestructive eigenvalue measurement for X-type stabilizers of the code CSS$(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$, that is, the Pauli operators $X(f^{k+1}),\ldots,X(f^m)$, where $f^{k+1},\ldots,f^m$ are the rows of G_0. If at least one of the measurement returns the outcome -1, the subroutine returns FAILED, and the final state is discarded. If all measured eigenvalues are $+1$, state η_2 has been projected onto the code space of the code CSS$(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$, and the subroutine is deemed successful (since we do not have any X-type errors, the syndrome of all Z-type stabilizers is automatically trivial). This results in a state

$$\eta_3 = \Pi_0\eta_2\Pi_0/P_s,$$

where Π_0 is the projector onto the code space of CSS$(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$ and $P_s = \mathrm{Tr}\,(\eta_2\Pi_0)$ is the success probability. State η_3 only has a contribution from errors $Z(f)$ with $f\in\mathcal{G}_0^\perp = \mathcal{G}_1\oplus\mathcal{G}^\perp$; see Lemma 1 since these are the only Z-type errors commuting with all X-type stabilizers. Hence the success probability is

$$P_s = \sum_{f\in\mathcal{G}_0^\perp}(1-p)^{n-|f|}p^{|f|} = \frac{1}{|\mathcal{G}_0|}\sum_{f\in\mathcal{G}_0}(1-2p)^{|f|}, \tag{14}$$

where the second equality uses the MacWilliams identity [24]. Any vector $f\in\mathcal{G}_1\oplus\mathcal{G}^\perp$ can be written as $f = g + x_1f^1 + \ldots + x_kf^k$, where $g\in\mathcal{G}^\perp$ and $f^1,\ldots,f^k$ are the rows of G_1. Since $Z(g)$ is a stabilizer, we conclude that

$$\begin{aligned} Z(f)|\overline{A^{\otimes k}}\rangle &= Z(x_1f^1+\cdots+x_kf^k)|\overline{A^{\otimes k}}\rangle \\ &= \overline{Z}_1^{x_1}\cdots\overline{Z}_k^{x_k}|\overline{A^{\otimes k}}\rangle. \end{aligned}$$

Here we used the definition of the logical Z-type operators; see Eq. (11). Hence state η_3 coincides with an encoded k-qubit mixed state

$$\rho_{\mathrm{out}} = \frac{1}{P_s}\sum_{x\in\mathbb{F}_2^k} p_{\mathrm{out}}(x)\,|A_x\rangle\,\langle A_x|\,, \tag{15}$$

where $|A_x\rangle = |A_{x_1}\rangle\otimes\cdots\otimes|A_{x_k}\rangle$ and

$$p_{\mathrm{out}}(x) = \sum_{f\in\mathcal{G}^\perp+x_1f^1+\cdots+x_kf^k}(1-p)^{n-|f|}p^{|f|}. \tag{16}$$

The last step of the subroutine is to decode CSS$(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$, thereby mapping η_3 to ρ_{out}. The k-qubit state ρ_{out} is the output state of the distillation subroutine. The reduced density matrix describing the ath output qubit can be written as

$$\rho_{\mathrm{out},a} = (1-q_a)\,|A_0\rangle\,\langle A_0| + q_a\,|A_1\rangle\,\langle A_1|\,,$$

where q_a is the output error rate on the ath qubit:

$$q_a = 1 - \frac{1}{P_s}\sum_{x\,:\,x_a=0} p_{\mathrm{out}}(x).$$

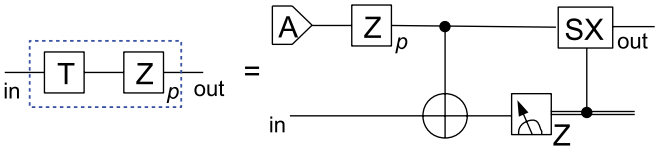

FIG. 1. (Color online) Implementation of the T gate using CO and one copy of the ancillary state $|A\rangle$. If the ancilla is a mixture of $|A\rangle$ and $Z\,|A\rangle$ with probabilities $1-p$ and p, respectively, the circuit enacts a noisy version of the T gate, namely, $\rho_{\mathrm{out}} = (1-p)T\rho_{\mathrm{in}}T^\dagger + pZT\rho_{\mathrm{in}}T^\dagger Z = \mathcal{E}\circ\mathcal{T}(\rho_{\mathrm{in}})$. The above circuit is used n times in the subroutine of Fig. 2.

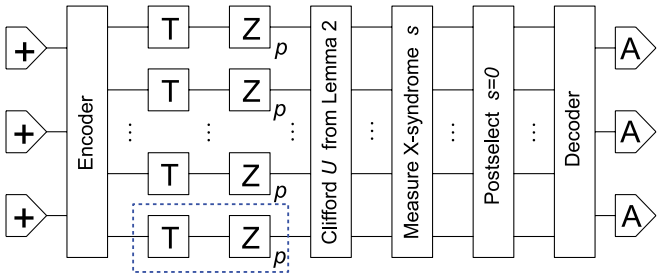

FIG. 2. (Color online) The distillation subroutine for the magic state $|A\rangle$ based on a triorthogonal matrix G. The encoder prepares k copies of the state $|+\rangle$ encoded by the stabilizer code CSS$(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$. Implementation of each T gate consumes one ancillary $|A\rangle$ state, as shown in Fig. 1. If the ancillae $|A\rangle$ have error rate p, each ideal T gate is followed by a Z error with probability p. The Clifford operator U is constructed in Lemma 2. Note that U is diagonal in the Z basis and thus commutes with any Z error. The syndrome s is measured only for X-type stabilizers $X(f^a)$, where f^a are the rows of G_0. In the case when all stabilizers $X(f^a)$ have eigenvalue $+1$ (trivial syndrome) the decoder is applied. It returns k copies of state $|A\rangle$ with the overall error probability $O(p^d)$. The trivial syndrome is observed with probability $1-O(p)$.

Let $\mathcal{K}_a$ be the sum of $\mathcal{G}^\perp$ and the space spanned by all rows of G_1 except for a. Lemma 1 implies that $\dim \mathcal{K}_a = \dim \mathcal{G}_0^\perp - 1$. On the other hand, $\mathcal{K}_a \subseteq (\mathcal{G}_0 \oplus (f^a))^\perp$, where $(f^a) = \{0^n, f^a\}$ is the one-dimensional subspace spanned by f^a. Hence $\mathcal{K}_a = (\mathcal{G}_0 \oplus (f^a))^\perp$, and thus

$$q_a = 1 - \frac{\sum_{f\in(\mathcal{G}_0\oplus(f^a))^\perp}(1-p)^{n-|f|}p^{|f|}}{\sum_{f\in\mathcal{G}_0^\perp}(1-p)^{n-|f|}p^{|f|}}. \tag{17}$$

We shall be mostly interested in the worst-case output error rate

$$q = \max_{a=1,\ldots,k} q_a. \tag{18}$$

Output qubits with $q_a < q$ can be additionally dephased in the A basis to achieve $q_a = q$. From Eq. (17) we infer that $q = O(p^d)$, where d is the minimum weight of a vector $f \in \mathcal{G}_0^\perp$ such that $(f, f^a) = 1$ for some a. Equivalently,

$$d = \min_{f\in\mathcal{G}_0^\perp\setminus\mathcal{G}^\perp} |f| \tag{19}$$

is the distance of the code $\mathrm{CSS}(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$ against Z-type errors. Using the MacWilliams identity, we also get

$$q_a = 1 - \frac{1}{2}\frac{\sum_{f\in\mathcal{G}_0\oplus(f^a)}(1-2p)^{|f|}}{\sum_{f\in\mathcal{G}_0}(1-2p)^{|f|}}. \tag{20}$$

This expression can be easily evaluated numerically in the important case when G_0 has only a few rows.

The above subroutine requires n extra qubits to prepare the encoded $|+^{\otimes k}\rangle$ state, while the total number of Pauli measurements is $n+m-k$. In Appendix A we describe an alternative subroutine which is slightly less intuitive but does not require any extra qubits and uses only $n-k$ Pauli measurements. Both subroutines output the same state and have the same success probability.

VI. FULL DISTILLATION PROTOCOL

The final goal of the distillation is to prepare a state σ of N qubits such that the overlap between σ and N copies of the magic state $|A\rangle$ is sufficiently close to 1, say, at least $2/3$. Such a state σ can be used as a resource to simulate any quantum circuit that contains Clifford gates and at most N gates of type T using only CO with an overall error probability of at most $1/3$. Each qubit of σ allows one to simulate one T gate using the scheme shown in Fig. 1.

Let σ_j be the reduced density matrix describing the jth qubit of σ. For any given target error rate ϵ our full protocol will distill a state σ which is diagonal in the basis $\{|A_0\rangle, |A_1\rangle\}^n$ and such that

$$\max_j \langle A_1|\sigma_j|A_1\rangle \leqslant \epsilon. \tag{21}$$

The standard union bound then implies that the overlap $\langle A_0^{\otimes N}|\sigma|A_0^{\otimes N}\rangle$ is close to 1 whenever $\epsilon \sim 1/N$.

In order to distill N magic states with the target error rate ϵ, the elementary subroutine described in Sec. V will be applied recursively such that each input state ρ consumed by a level m distillation subroutine is one of the output states $\rho_{\mathrm{out},a}$ distilled by some level $(m-1)$ subroutine. The recursion starts at a level $m=0$ with NC input states, where $C = C(\epsilon)$ is the distillation cost. In the limit $N \gg 1$ the distillation rounds can be organized such that all n input states ρ consumed by any elementary subroutine at a level m have been distilled at *different* subroutines at the level $m-1$; see Lemma IV in [18]. This allows one to disregard correlations between errors and analyze the full protocol using the average *yield*

$$\Gamma(p) = \frac{kP_s(p)}{n},$$

that is, the average number of output states with an error rate $q(p)$ per one input state with an error rate p. Here q is defined in Eqs. (18) and (20). Neglecting the fluctuations, the distillation cost C, the input error rate p, the target error rate ϵ, and the required number of levels m_0 are related by the following obvious equations:

$$\begin{aligned} C_{m+1} &= \Gamma(p_m)C_m, \\ p_{m+1} &= q(p_m), \quad m = 0,\ldots, \quad m_0-1, \\ p_{m_0} &= \epsilon, \quad p_0 = p, \quad C_{m_0} = 1, \quad C_0 = C. \end{aligned} \tag{22}$$

In the limit of small p one has $P_s(p) \approx 1$ and thus $\Gamma(p) \approx k/n$. Taking into account that $q = O(p^d)$, where the distance d is defined in Eq. (19), one arrives at

$$C(\epsilon) = O(\log^\gamma(1/\epsilon)), \quad \gamma = \frac{\log(n/k)}{\log(d)}, \tag{23}$$

provided that the input error rate p is below a constant threshold value p_{th}, which depends on the chosen triorthogonal matrix.

We conjecture that the scaling exponent γ of the distillation cost C cannot be smaller than 1 for any concatenated distillation protocol based on a triorthogonal matrix. Indeed, suppose the output error rate satisfies $q(p) \leqslant cp^d < p$ for $p < p_0$ and $q(1) = 1$. As noted above, the potential correlation in the error probabilities among the output states may be ignored. Then, after m levels of distillation the output error rate should satisfy

$$\epsilon \leqslant c^{-1/(d-1)}(c'p_0)^{d^m},$$

where $c' = c^{(2-d)/(d-1)}$. Let $\alpha = n/k$ be the inverse yield in the small input error rate limit. Clearly, $C \geqslant \alpha^m$. Since $q(1) = 1$, the probability that the output is the desired magic state can be at most $1 - p_0^C$. It follows that $p_0^C \leqslant \epsilon$, and therefore $\alpha \geqslant d$. We conclude that

$$C \geqslant d^m = \Omega(\log(1/\epsilon)).$$

VII. A FAMILY OF TRIORTHOGONAL MATRICES

To construct explicit distillation protocols, triorthogonal matrices G with high yield k/n are called for. A natural strategy to maximize the yield is to keep the number of even-weight rows in G as small as possible. Indeed, each extra row in G_0 increases the number of constraints due to Eqs. (1) and (2) without increasing the yield. However, the number of rows in G_0 cannot be too small. Recall that the distillation subroutine of Sec. V improves the quality of magic states only if $d \geqslant 2$, where d is the distance of the code $\mathrm{CSS}(X,\mathcal{G}_0; Z,\mathcal{G}^\perp)$ against Z errors defined in Eq. (19). We claim that $d = 1$ whenever G_0 has fewer than three rows. Indeed, in this case Lemma 3 implies that G_0 must have a zero column, say, the first one. Then $e_1 \equiv$

$(1,0,\ldots,0) \in \mathcal{G}_0^{\perp}$. On the other hand, $e_1 \notin \mathcal{G}^{\perp}$ since otherwise the first column of G would be zero. This shows that $d = 1$ [see Eq. (19)]. Hence a good strategy is to look for candidate triorthogonal matrices with three even-weight rows such that G_0 has no zero columns. This guarantees $d \geqslant 2$.

Below we present a family of triorthogonal matrices with yield $k/n = k/(3k+8)$, where k is even. The matrices are constructed from several simple submatrices, which we define as:

$$L = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$
$$S_1 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{24}$$

For each even number $k \geqslant 0$, define the $(k+3) \times (3k+8)$ matrix

$$G(k) = \begin{bmatrix} 0 & L & M & 0 & \cdots & 0 \\ 0 & L & 0 & M & & 0 \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ 0 & L & 0 & 0 & \cdots & M \\ S_1 & S_1 & S_2 & S_2 & \cdots & S_2 \end{bmatrix}, \tag{25}$$

where L, M, and S_2 appear $k/2$ times.

This family of matrices is triorthogonal, with k odd-weight rows and three even-weight rows. To see this, first consider the usual orthogonality condition (1). Any pair of rows from $G(k)_1$, the upper k rows, overlap in L, which has weight 4. The bottom three rows, $G(k)_0$, give three pairs whose overlaps have weights 4,4, and $2+k$, respectively. A row from $G(k)_1$ and another from $G(k)_0$ overlap at four positions. Thus the rows of $G(k)$ are mutually orthogonal. One can similarly check the triorthogonality condition (2).

For any linear space $\mathcal{F} \subseteq \mathbb{F}_2^n$ define its weight enumerator as $W_{\mathcal{F}}(x) = \sum_{f \in \mathcal{F}} x^{|f|}$. The error analysis in Sec. V requires the weight enumerators of $\mathcal{G}(k)_0$ and $\mathcal{G}(k)_0 \oplus (g^a)$ for all $a = 1,\ldots,k$, where g^a are the rows of $G(k)_1$. Due to the periodic structure of $G(k)$, the weight enumerator of $\mathcal{G}(k)_0 \oplus (g^a)$ is independent of a. The classical codes $\mathcal{G}(k)_0$ and $\mathcal{G}(k)_0 \oplus (g^1)$ have only 8 and 16 code vectors, respectively, and therefore an explicit calculation is easy:

$$\begin{aligned} W_{\mathcal{G}(k)_0}(x) &= 1 + x^8 + 6x^{4+2k}, \\ W_{\mathcal{G}(k)_0 \oplus (g^1)}(x) &= 1 + 2x^7 + x^8 + 6x^{3+2k} + 6x^{4+2k}. \end{aligned} \tag{26}$$

If $G(k)$ is used in our distillation protocol, the success probability or acceptance rate given the input error rate p is

$$P_s(p) = 1 - (8+3k)p + \cdots,$$

and the output error rate q on any *one* qubit is

$$q(p) = (1+3k)p^2 + \cdots$$

using Eq. (20), where $\cdots$ indicate higher-order terms in p. The initial term of $q(p)$ can be intuitively understood. Since the stabilizer code $\mathrm{CSS}(X, \mathcal{G}(k)_0; Z, \mathcal{G}(k)^{\perp})$ has logical Z operators of weight 2, the probability that there is an undetected error on the output qubit is $O(p^2)$. The coefficient of p^2 is the number of logical Z operators of weight 2 that acts nontrivially on a particular logical qubit, which is readily counted as $4 + 3(k-1)$.

The threshold input error rate can be obtained by the requirement that $q(p) < p$. From the leading term of $q(p)$, one may estimate the threshold as

$$p_{\mathrm{th}} \approx \frac{1}{3k+1}.$$

Provided that the input error rate is smaller than p_{th}, solving Eq. (22) gives

$$C(\epsilon) = O\left(\log^{\gamma} \frac{1}{\epsilon}\right), \quad \gamma = \log_2 \frac{3k+8}{k}.$$

The scaling exponent γ reaches $\log_2 3 \approx 1.585$ in the large k limit, which is the best to the authors' awareness.

VIII. COMPARISON WITH KNOWN PROTOCOLS

The output error rate improves most when the input error rate is much smaller than the threshold of the protocol.

TABLE I. Minimum average number C of required input magic states of the fixed error rate $p_{\mathrm{in}} = 0.01$ to distill a single-output magic state of error rate $\leqslant \epsilon_{\mathrm{target}}$. The sequence of labels in the second column denotes the subroutines in order from left to right. An even number k in the second column denotes the one round of distillation using $G(k)$. Here "15" and "5" respectively represent the protocols by [13] and [18]. C_{MEK} utilized only 15 and 5. This table is numerically optimized under the restriction that there be at most five rounds of distillation.

$-\log_{10} \epsilon_{\mathrm{target}}$	Protocol	$-\log_{10} \epsilon_{\mathrm{actual}}$	C	C_{MEK}
3	5	3.030	5.521	5.521
4	15	4.443	17.44	17.44
5	5-5	5.104	27.86	27.86
6	15-40	6.802	56.07	83.99
7	15-24	7.022	58.30	83.99
8	5-5-40	8.125	89.26	139.3
9	5-5-5	9.253	139.3	139.3
10	15-40-40	11.52	179.4	261.7
11	15-40-40	11.52	179.4	261.7
12	15-24-36	12.01	187.9	418.0
13	15-10-20	13.00	225.6	418.0
14	5-5-40-40	14.17	285.6	419.9
15	5-5-18-28	15.00	315.5	696.7
16	5-5-6-22	16.03	406.2	696.7
17	5-5-5-10	17.02	529.5	696.7
18	15-40-40-40	20.96	574.1	1260
19	15-40-40-40	20.96	574.1	1260
20	15-40-40-40	20.96	574.1	1260
21	15-38-40-40	21.05	575.9	1260
22	15-22-38-40	22.03	604.3	1308
23	15-14-30-40	23.01	652.3	2090
24	15-10-18-40	24.01	731.5	2090
25	15-6-16-36	25.01	853.1	2090
26	5-5-40-40-40	26.25	914.0	2090
27	5-5-26-38-40	27.04	947.5	2100
28	5-5-16-32-40	28.01	1015	2181
29	5-5-10-26-38	29.01	1125	3483
30	5-5-8-14-30	30.01	1301	3483

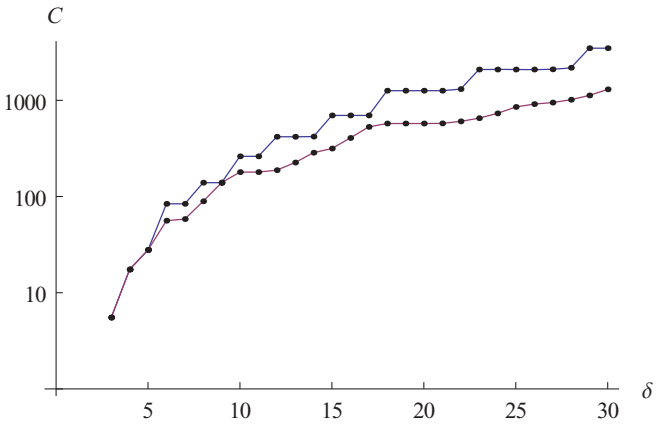

FIG. 3. (Color online) Distillation cost C as a function of the target error rate $\epsilon = 10^{-\delta}$ for a fixed input error rate $p = 0.01$. The top curve is obtained from [18], and the bottom curve is obtained from the optimization using the triorthogonal matrices $G(k)$. The lines are a guide to the eye.

One cannot thus use $G(k)$ naively with large k since the threshold is inversely proportional to k. It is therefore desirable to concatenate various protocols to minimize the resource requirement. This optimization is carried out for illustrative purpose by a numerical computation. We restrict the number of rounds to be less than or equal to 5 and consider all possible combinations of (i) "15," the 15-to-1 protocol [13], (ii) "5," the 10-to-2 protocol [18], (iii) "k," the $(3k+8)$-to-k protocol using the triorthogonal matrices $G(k)$ for $k = 2,4,6,\ldots,40$, and (iv) "49," the 49-to-1 protocol presented in Appendix B. The result is summarized in Table I, where the numbers in the quotation marks above are used to denote each subroutine. Unfortunately, the 49-to-1 protocol has found no place in the best combinations. See also Fig. 3. A general rule is that it is better to use high-threshold protocols for initial rounds and then use high-yield protocols when the error rate becomes small.

IX. LINEAR EQUATIONS FOR TRIORTHOGONAL MATRICES

The triorthogonality equations (1) and (2) in general depend on a particular presentation of G and are not automatically guaranteed by the classical code $\mathcal{G}$. However, a certain choice of variables associated with G yields a set of linear equations over $\mathbb{F}_2$, equivalent to the triorthogonality. This system of linear equations makes a numerical search effective.

Suppose a triorthogonal matrix G is of size $m \times n$. Let $x = (x_1,\ldots,x_m) \in \mathbb{F}_2^m$ denote an arbitrary m-bit string. Each column of the matrix G corresponds to a particular $x \in \mathbb{F}_2^m$; in other words, G is described by n such bit strings x. The cardinality of the overlap between the ath and bth rows $(a \neq b)$ is exactly the number of columns x in G such that $x_a = x_b = 1$. Let N_x be the number of columns x appearing in G. Then, the usual orthogonality condition (1) can be written as

$$\sum_{x\in\mathbb{F}_2^m : x_a = x_b = 1} N_x = 0 \pmod 2 \tag{27}$$

for distinct a,b. Likewise, the cardinality of the triple overlap among distinct rows a,b,c is exactly the number of columns x such that $x_a = x_b = x_c = 1$. Therefore, the triorthogonality condition (2) is equivalent to

$$\sum_{x\in\mathbb{F}_2^m : x_a = x_b = x_c = 1} N_x = 0 \pmod 2 \tag{28}$$

for distinct a,b,c. The weight of each row a is the sum $\sum_{x:x_a=1} N_x$. Demanding k odd-weight rows of G is possible with the following inhomogeneous equations:

$$\sum_{x\in\mathbb{F}_2^m : x_a = 1} N_x = \begin{cases} 1 \pmod 2 & \text{if} \quad 1 \leqslant a \leqslant k, \\ 0 \pmod 2 & \text{otherwise.} \end{cases} \tag{29}$$

Conversely, treating all N_x as unknown binary variables, any solution to Eqs. (27)–(29) gives rise to a triorthogonal matrix. Namely, we just write a column $x^T = (x_1,\ldots,x_m)^T$ whenever $N_x = 1$. The number of columns of the resulting matrix will be the Hamming weight of the vector N, whose components are indexed by $x \in \mathbb{F}_2^m$.

One does not have to be concerned about the situation $N_x > 1$ because it only produces less efficient protocols for magic-state distillation. Suppose there are repeated columns in an $n' \times m$ triorthogonal matrix G', and let G be the $n \times m$ triorthogonal matrix obtained from G' by removing repeated columns in pairs. Consider $Z(f)$, a logical operator of $\mathcal{C}' = \mathrm{CSS}(X,\mathcal{G}_0'; Z,\mathcal{G}'^{\perp})$ of minimal weight. The support of f should not involve any pair of indices of the repeated columns due to the minimality. Hence, $Z(f)$ may be thought of as a logical operator of $\mathcal{C} = \mathrm{CSS}(X,\mathcal{G}_0; Z,\mathcal{G}^{\perp})$. Conversely, any logical operator of $\mathcal{C}$ can be viewed as that of $\mathcal{C}'$. Therefore, $\mathcal{C}$ and $\mathcal{C}'$ have the same minimal weight for Z-type logical operators, but $\mathcal{C}'$ has longer length. For the same reason, it is safe to assume $N_{(0,0,\ldots,0)} = 0$.

The set of all solutions to Eqs. (27)–(29) contains useless triorthogonal matrices. In order for a protocol to be useful, the minimal weight for Z-type logical operators must be at least 2. If a triorthogonal matrix G has an all-zero column in G_0, the lower $m-k$ even-weight rows, then the resulting stabilizer code $\mathrm{CSS}(X,\mathcal{G}_0; Z,\mathcal{G}^{\perp})$ admits a weight-one Z-type logical operator. Thus, we should impose the following linear constraints:

$$N_{(x_1,\ldots,x_k,0,\ldots,0)} = 0 \tag{30}$$

for all $(x_1,\ldots,x_k) \in \mathbb{F}_2^k$.

So, given the number m of rows of G and the number k of odd-weight rows, one can solve the above equations over $\mathbb{F}_2$ to find the minimal weight solution N. There are 2^m variables N_x and $2^k + \binom{m}{1} + \binom{m}{2} + \binom{m}{3}$ equations. Note that due to Lemma 3, one has to consider the case $m - k \geqslant 3$.

ACKNOWLEDGMENTS

J.H. is in part supported by the Institute for Quantum Information and Matter (IQIM), an NSF Physics Frontier Center, and by the Korea Foundation for Advanced Studies. J.H. is grateful for the hospitality of the IBM Watson Research Center, in the form of a summer internship while this work is done. S.B. was partially supported by the DARPA QUEST program under Contract No. HR0011-09-C-0047 and by the

Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center Contract No. D11PC20167.

APPENDIX A: ALTERNATIVE DISTILLATION SUBROUTINE

In this appendix we show that the distillation scheme proposed in Ref. [13] can be adapted to any stabilizer code based on a triorthogonal matrix. It can serve as an alternative to the subroutine described in Sec. V. Both subroutines output the same state and have the same success probability.

Let G be any triorthogonal matrix with n columns, k odd-weight rows $f^1,\ldots,f^k$, and $m-k$ even-weight rows. Consider the following distillation protocol that takes n input qubits and outputs k qubits.

(1) Measure eigenvalues of $Z(f)$, $f\in\mathcal{G}^\perp$. Let the eigenvalue of $Z(f)$ be $(-1)^{\mu(f)}$, where $\mu:\mathcal{G}^\perp\to\mathbb{F}_2$ is a linear function (Z syndrome).

(2) Choose any $w\in\mathbb{F}_2^n$ such that $\mu(f)=(w,f)$ for all $f\in\mathcal{G}^\perp$. Apply $A(w)^\dagger$.

(3) Apply unitary U from Lemma 2.

(4) Measure eigenvalues of $X(g)$, $g\in\mathcal{G}_0$. Declare FAILED unless all eigenvalues are $+1$.

(5) Decode $\mathrm{CSS}(X,\mathcal{G}_0;Z,\mathcal{G}^\perp)$.

Note that the measurements of $Z(f)$ and $X(g)$ in steps 1 and 4 only need to be performed for basis vectors $f\in\mathcal{G}^\perp$ and $g\in\mathcal{G}_0$, respectively. Hence the total number of Pauli measurements is

$$\dim(\mathcal{G}^\perp)+\dim(\mathcal{G}_0)=(n-m)+(m-k)=n-k.$$

Let $\rho=(1-p)\,|A_0\rangle\langle A_0|+p\,|A_1\rangle\langle A_1|$ be the raw ancilla. We claim that the above protocol maps $\rho^{\otimes n}$ to the output state defined in Eqs. (15) and (16), while the success probability $P_s(p)$ is given by Eq. (14). Indeed, since the input state $\rho^{\otimes n}$ is diagonal in the A basis and the correcting operator $A(w)^\dagger$ has the same Z syndrome as the one measured at step 1, the state obtained after step 2 is

$$\eta_2=\Pi_Z\rho^{\otimes n}\Pi_Z/\mathcal{Z},$$

where Π_Z projects onto the subspace with the trivial Z syndrome and $\mathcal{Z}$ is a normalizing coefficient such that $\mathrm{Tr}(\eta_2)=1$. Since $\rho=\mathcal{E}(|A\rangle\langle A|)$, where $\mathcal{E}$ involves only Z errors [see Eq. (13)], one gets

$$\mathcal{Z}=\langle A^{\otimes n}|\Pi_Z|A^{\otimes n}\rangle=\langle +^n|\Pi_Z|+^n\rangle. \tag{A1}$$

Consider a pair of codes

$$\mathcal{C}_X\equiv\mathrm{CSS}(X,\mathcal{G}_0;Z,\mathcal{G}^\perp),\quad \mathcal{C}_A\equiv\mathrm{CSS}(A,\mathcal{G}_0;Z,\mathcal{G}^\perp),$$

where we adopt the notation of Ref. [13]. Note that $\mathcal{C}_A$ has non-Pauli stabilizers $A(g)$, $g\in\mathcal{G}_0$, in addition to Pauli ones $Z(g)$, $g\in\mathcal{G}^\perp$. By abusing the notation we shall sometimes identify $\mathcal{C}_X$ and $\mathcal{C}_A$ with the code spaces of the respective codes. Taking into account that $A=TXT^\dagger$ and $TZ=ZT$, we conclude that $\mathcal{C}_A=\hat{T}\cdot\mathcal{C}_X$, where $\hat{T}=T^{\otimes n}$. Let U be the diagonal Clifford unitary constructed in Lemma 2. From Eq. (10) we infer that $U\hat{T}$ preserves the code space $\mathcal{C}_X$ and thus

$$U\cdot\mathcal{C}_A=\mathcal{C}_X. \tag{A2}$$

This shows that $|\psi\rangle\in\mathcal{C}_A$ can be specified by eigenvalue equations $\Pi_Z|\psi\rangle=|\psi\rangle$ and

$$U^\dagger X(g)U|\psi\rangle=|\psi\rangle\quad\text{for all}\quad g\in\mathcal{G}_0. \tag{A3}$$

To analyze the rest of the protocol it will be convenient to insert two dummy steps between steps 4 and 5, namely, step 4a, apply $U^\dagger$, and step 4b, apply U. Taking into account Eq. (A3), we conclude that the overall effect of steps 1–4a is to project the state $\rho^{\otimes n}$ onto the code space $\mathcal{C}_A$. Let Π_A be the projector onto the subspace with the trivial A syndrome of the code $\mathcal{C}_A$. Then the (unnormalized) state obtained after step 4a is

$$\eta_{4a}=\Pi_Z\Pi_A\rho^{\otimes n}\Pi_A\Pi_Z/\mathcal{Z},$$

while the success probability is determined by $P_s=\mathrm{Tr}(\eta_{4a})$. Consider any term $\Pi_AZ(f)|A^{\otimes n}\rangle$ in η_{4a}. Since $\Pi_A|A^{\otimes n}\rangle=|A^{\otimes n}\rangle$, the state η_{4a} gets contributions only from errors $Z(f)$ such that $\Pi_AZ(f)\Pi_A\neq 0$. Such errors must commute with any A-type stabilizer, which is possible only if $f\in\mathcal{G}_0^\perp$. In this case one has $\Pi_AZ(f)=Z(f)\Pi_A$. This shows that

$$\eta_{4a}=\frac{1}{\mathcal{Z}}\tilde{\mathcal{E}}(\Pi_Z|A\rangle\langle A|^{\otimes n}\Pi_Z),$$

where $\tilde{\mathcal{E}}$ is a linear map defined as

$$\tilde{\mathcal{E}}(\eta)=\sum_{f\in\mathcal{G}_0^\perp}(1-p)^{n-|f|}p^{|f|}Z(f)\eta Z(f).$$

The identity $|A\rangle=T|+\rangle$ and Lemma 2 yield

$$\frac{\Pi_Z|A^{\otimes n}\rangle}{\sqrt{\mathcal{Z}}}=\frac{\hat{T}\Pi_Z|+^{\otimes n}\rangle}{\sqrt{\mathcal{Z}}}=\hat{T}\;|G\rangle=U^\dagger|\overline{A^{\otimes k}}\rangle.$$

Note that all states above are normalized. Thus the state obtained after step 4b (i.e., after step 4 of the original protocol) is

$$\eta_4=\tilde{\mathcal{E}}(|\overline{A^{\otimes k}}\rangle\langle\overline{A^{\otimes k}}|).$$

This shows that $P_s=\mathrm{Tr}(\eta_4)$ is indeed given by Eq. (14). As was shown in Sec. V, decoding state η_4 yields the desired output state, Eq. (15).

APPENDIX B: THE 49-TO-1 PROTOCOL

The approach pursued in this paper aims to minimize the distillation cost scaling exponent $\gamma=\log_2(n/k)/\log_2 d$ by constructing codes with high yield k/n and $d=2$. An alternative method of constructing codes with large distance d and small yield (e.g., $k=1$) appears to be less fruitful. Using the linear system method of Sec. IX, we were able to find a 49-qubit code with $k=1$ that admits a transversal T gate and has distance $d=5$. The corresponding triorthogonal matrix

G_0 of size 13×49 is

$$G_0 = \begin{bmatrix}
1111111111111101010101010101010101010101010101 \\
0000000000000000001111001100110000110011001100110011 \\
0000000000000000110000001100110011000000000000000000 \\
0000000000000000000000000000000000001111000000001111 \\
0000000000000000011110000000000000000011110000000000 \\
0000000000000000000001111000011110000000000000000000 \\
0000000000000000000000001111111100000000000000000000 \\
0000000000000000000000000000000011111111000000000000 \\
0000000000000000000000000000000000000000000011111111 \\
1010101010101010000000000000000000000000000000000000 \\
0110011001100110000000000000000000000000000000000000 \\
0001111000011110000000000000000000000000000000000000 \\
0000000111111111000000000000000000000000000000000000
\end{bmatrix}.$$

The weight enumerator of $\mathcal{G}_0$ computed numerically is

$$W_{49}(x) = 1 + 32x^8 + 442x^{16} + 6696x^{24} + 1021x^{32}.$$

Thus, $\mathcal{G}_0$ is a triply even linear code [25]; that is, $|f| = 0 \pmod 8$ for any $f \in \mathcal{G}_0$. By adding an all-ones row to G_0, one obtains a triorthogonal matrix G with $k = 1$. It leads to a protocol distilling one magic state out of 49 input states. Note that for any triorthogonal matrix with one odd-weight row 1^n the relevant distance d defined in Eq. (19) can be written as

$$d = \min_{\substack{f \in \mathcal{G}_0^\perp \\ |f| \text{ is odd}}} |f|. \tag{B1}$$

We have checked numerically that $d = 5$ for the 49-qubit code. Since the code is triply even, the Clifford operator U defined in Lemma 2 is the identity. The output error rate as a function of input error rate has the leading term

$$q_{49}(p) = 1411p^5 + \cdots.$$

The distillation threshold was found to be $p_{49;\text{th}} = 0.1366$.

We note that the above 49-qubit code is optimal in the sense that there are no triply even linear codes of odd length $n \leqslant 47$ such that the distance d defined in Eq. (B1) is greater than 3. This fact can be checked numerically using the classification of all maximal triply even codes of length 48 found in [25]. A maximal triply even code of length 47 or shorter can be thought of as a subcode of some maximal triply even code of length 48 obtained by imposing the linear condition for one component to be zero. Using the results of [25], we were able to examine numerically all maximal triply even codes of length 47. We found that $d \leqslant 3$ for all such codes. Further shortening cannot increase the distance d.

[1] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, J. Math. Phys. **43**, 4452 (2002).

[2] P. W. Shor, in *Proceedings of 37th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, USA, 1996), pp. 56–65.

[3] E. Knill, arXiv:quant-ph/0404104.

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

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

[6] R. Raussendorf and J. Harrington, Phys. Rev. Lett. **98**, 190504 (2007).

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

[8] R. Raussendorf, J. Harrington, and K. Goyal, New J. Phys. **9**, 199 (2007).

[9] D. Gottesman, arXiv:quant-ph/9705052.

[10] H. Bombin and M. A. Martin-Delgado, J. Phys. A **42**, 095302 (2009).

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

[12] S. Bravyi and R. Koenig, arXiv:1206.1609.

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

[14] G. Nebe, E. M. Rains, and N. J. A. Sloane, arXiv:math/0001038.

[15] Notation used in the present paper is slightly different from that of Ref. [13]. In particular, the $\pi/8$ rotation T should not be confused with the Clifford gate denoted by T in Ref. [13].

[16] P. O. Boykin, T. Mor, M. Pulver, V. P. Roychowdhury, and F. Vatan, Inf. Process. Lett. **75**, 101 (2000).

[17] B. W. Reichardt, Quantum Inf. Process. **4**, 251 (2005).

[18] A. M. Meier, B. Eastin, and E. Knill, arXiv:1204.4221.

[19] E. T. Campbell, H. Anwar, and D. E. Browne, arXiv:1205.3104.

[20] T. Jochym-O'Connor, Y. Yu, B. Helou, and R. Laflamme, arXiv:1205.6715.

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

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

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

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

[25] K. Betsumiya and A. Munemasa, J. London Math. Soc. **86**, 1 (2012).