Magic State Distillation with Low Space Overhead and Optimal Asymptotic Input Count

Jeongwan Haah[1], Matthew B. Hastings[2,1], D. Poulin[3], and D. Wecker[1]

[1]Quantum Architectures and Computation, Microsoft Research, Redmond, WA 98052, USA
[2]Station Q, Microsoft Research, Santa Barbara, CA 93106-6105, USA
[3]Département de Physique & Institut Quantique, Université de Sherbrooke, Quebec, Canada

26 September 2017

We present an infinite family of protocols to distill magic states for T-gates that has a low space overhead and uses an asymptotic number of input magic states to achieve a given target error that is conjectured to be optimal. The space overhead, defined as the ratio between the physical qubits to the number of output magic states, is asymptotically constant, while both the number of input magic states used per output state and the T-gate depth of the circuit scale linearly in the logarithm of the target error δ (up to $\log\log 1/\delta$). Unlike other distillation protocols, this protocol achieves this performance without concatenation and the input magic states are injected at various steps in the circuit rather than all at the start of the circuit. The protocol can be modified to distill magic states for other gates at the third level of the Clifford hierarchy, with the same asymptotic performance. The protocol relies on the construction of weakly self-dual CSS codes with many logical qubits and large distance, allowing us to implement control-SWAPs on multiple qubits. We call this code the "inner code". The control-SWAPs are then used to measure properties of the magic state and detect errors, using another code that we call the "outer code". Alternatively, we use weakly-self dual CSS codes which implement controlled Hadamards for the inner code, reducing circuit depth. We present several specific small examples of this protocol.

The possibility of a large scale quantum computer relies on fault-tolerant architectures, in which errors are corrected faster than they are created [1–3]. The standard approach is to use stabilizer codes to protect logical qubits from noise [4, 5], and perform quantum gates at the encoded level. The overhead of the fault-tolerance is *only* polynomial in the logarithm of the desired accuracy, but in practice the overhead is estimated to be overwhelmingly large [6, 7]. Particularly expensive operations are non-Clifford gates such as the $\pi/4$-rotation (T-gate) and Toffoli gate. A compelling approach is to inject a special state, called a magic state, into a Clifford-only circuit, and pass the cost of implementing the non-Clifford operation to the preparation of the magic states, which are distilled from noisy ones [8–10].

There exist several distillation protocols for the magic state (T-state) using specialized quantum error correcting codes [8, 10–12]. Each code can provide a fixed degree of fidelity

improvement that is given by the code distance. In order to achieve arbitrary good fidelity, one typically concatenates small routines. In terms of the number of input magic states of low fidelity per one output magic state of high fidelity, the best protocols to date are those in Refs. [12, 13]. However, these protocols require a large batch of thousands of magic states to be useful.

In this paper, we introduce an infinite family of distillation protocols, extending one of the very first protocols by Knill [8], and another by Meier, Eastin, and Knill [11]. Our protocol produces n T-magic states using at most cn qubits and achieves at least $c'n$-th order error suppression under the assumption that the sole noise source is the T gate, where c, c' are small universal constants. Since the degree of error suppression is high, there is no need to concatenate small routines, reducing the space overhead significantly.

Our protocol is also asymptotically superior (conjectured to be optimal) in terms of noisy T count. For any fixed odd $d \geq 5$, we show that the number of noisy T gates per one output magic state with error suppressed to d-th order converges to d exactly in the large code length limit.

Beyond the magic states for T gates, our protocol can distill magic states for rotation by $\pi/2^k$ for $k = 3, 4, \ldots$ adapting the idea of Ref. [14], and any gate in the third level of Clifford hierarchy [15]. (See also Ref. [16] for smaller angle ($k \geq 3$) rotations, though we do not use ideas there.) For the latter, the asymptotic performance is similar to the T gate case.

Small instances of our family demonstrates reduction of space overhead, with a modest input T count. If noisy $\pi/4$ rotations can be directly done on qubits, an instance of our family operates on 34 qubits including measurement ancillas, produces 15 T-magic states with 5th order error suppression, and requires 29 noisy T gates per output. In comparison, to the authors' knowledge, any previous protocol that can operate on less than 50 qubits have either have lower order of error suppression, or requires more non-Clifford gates per output.

Recent innovations show that the $\pi/4$-rotation and Toffoli gate can be implemented fault-tolerantly on a class of error correcting codes [17–25]. These schemes achieve computational universality through local operations while circumventing no-go theorems [26, 27] by going back and forth between two code spaces. This approach removes the need for magic states, but, it is not a simple question to tell which approach is better. This question depends on an architecture and underlying physical qubits' characteristic, and thus we leave the realistic cost analysis and comparison to future work.

The organization of the present paper is as follows. We start in Section 1 with a basic explanation of our ideas by exhibiting examples of small sizes. Section 2 explains how to convert magic state distillation protocols for T gates into those for Toffoli gates. In Section 3 we show that any weakly self-dual CSS code can be used in distillation protocols by implementing measurement of Clifford operators. In Section 4, we give asymptotic constructions of the codes, in the limit of either large distance or large code length. In Section 5, we give results of numerical simulations; in this section we also present some specific additional small size protocols that are not described elsewhere in the paper. We conclude with discussion in Section 6. In Appendix A, we give details, including stabilizer checks, for some of the specific codes used in the paper. Appendix B gives circuits for some of the protocols used. Appendix C describes unexpected relations among different distillation protocols. Appendix D explains an extension to qudits, using classification of symmetric forms over finite fields in Appendix E.

Throughout this paper, all classical codes that we consider will be linear codes over the binary field $\mathbb{F}_2$ and all quantum codes will be qubit stabilizer codes (except for Appendix D).

Given a bit vector v, we use $|v|$ to denote its Hamming weight. Our magic state is the $(+1)$-eigenstate $|H\rangle = \cos\frac{\pi}{8}|0\rangle + \sin\frac{\pi}{8}|1\rangle$ of the Hadamard operator H. We use matrices

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{0.1}$$

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad T = e^{-i\pi Y/8} = \begin{pmatrix} \cos\frac{\pi}{8} & -\sin\frac{\pi}{8} \\ \sin\frac{\pi}{8} & \cos\frac{\pi}{8} \end{pmatrix}. \tag{0.2}$$

A self-orthogonal subspace of some vector space is a subspace such that the inner product of any two vectors in the subspace vanishes. A weakly self-dual CSS code is a quantum code whose stabilizer group is invariant under conjugation by the product of the Hadamard operator on all qubits; we call this product a "transversal Hadamard".

1 Basic Distillation Protocols

Distillation protocols for magic states to date can be put roughly into three classes. Those in the first class implement a non-Clifford $\pi/4$-rotation to a stabilizer state such as $|+\rangle$ or $|0\rangle$ [10, 12, 16]. The non-Clifford rotation must be done fault-tolerantly, so the protocols in this class focus on finding error correcting codes that admits a transversal non-Clifford rotation. This requires the underlying code to have a special symmetry, which is rather rare. The protocols in the second class [8, 9, 11, 13, 14, 28] implement measurements of "stabilizers" of the magic state, based on the fact that a magic state in the third level of Clifford hierarchy [15] is an eigenstate of a Clifford operator. To measure a Clifford operator one needs a non-Clifford operation which has been implemented fault-tolerantly by a distance-two code. The third class uses yet different symmetries of codes [10, 29], and has high threshold for distillation, but the success probability does not reach 1 even with perfect input magic states.

Our scheme in the present paper belongs to the second class, and is an extension of the idea of Knill [8]. There are two levels of error correcting codes in our scheme, which we call *inner* and *outer* codes. Roughly speaking, the outer codes specify a certain set of measurements of Clifford operators on a set of input magic states, and the inner codes specify how to implement these measurements. We illustrate aspects of our ideas by two examples. They are not the best protocols in regards to e.g. the total number of non-Clifford gates and states, but will be simple to explain. A more general class of protocols is presented in later sections.

Without loss of generality, by a standard Clifford twirling argument, we can assume that each $\pi/4$ rotation and undistilled magic state suffers from independent Y errors with probability ϵ. We refer to this error model as the **stochastic** error model.

1.1 Trivial outer code

If we could implement the control-Hadamard, then the distillation is trivial: Prepare an ancilla qubit in $|+\rangle$ state, apply the control-Hadamard with the control on the ancilla and the Hadamard on an arbitrary target qubit, and measure the ancilla in X-basis to accept $+1$ outcome. The accepted target qubit is projected onto the magic state.

The control-Hadamard belongs to the third level of Clifford hierarchy, and thus cannot be implemented with Clifford operations. To obtain an approximate control-Hadamard with

noisy non-Clifford rotations, we must use an error correcting code that can implement H on the logical qubits fault-tolerantly.

1.1.1 $[[7,1,3]]$ inner code

To this end, we observe that the Steane code [30] admits a transversal Hadamard [8, Sec. IX]. The stabilizers are

$$\begin{array}{ccccccc} \mathrm{I} & \mathrm{I} & \mathrm{I} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} \\ \mathrm{I} & \mathrm{X} & \mathrm{X} & \mathrm{I} & \mathrm{I} & \mathrm{X} & \mathrm{X} \\ \mathrm{X} & \mathrm{I} & \mathrm{X} & \mathrm{I} & \mathrm{X} & \mathrm{I} & \mathrm{X} \end{array}, \quad \begin{array}{ccccccc} \mathrm{I} & \mathrm{I} & \mathrm{I} & \mathrm{Z} & \mathrm{Z} & \mathrm{Z} & \mathrm{Z} \\ \mathrm{I} & \mathrm{Z} & \mathrm{Z} & \mathrm{I} & \mathrm{I} & \mathrm{Z} & \mathrm{Z} \\ \mathrm{Z} & \mathrm{I} & \mathrm{Z} & \mathrm{I} & \mathrm{Z} & \mathrm{I} & \mathrm{Z} \end{array}, \tag{1.1}$$

the group generated by which is fixed under the $H^{\otimes 7} : X_i \leftrightarrow Z_i$, and the logical operator pair is

$$\mathrm{Z}\ \mathrm{Z}\ \mathrm{Z}\ \mathrm{Z}\ \mathrm{Z}\ \mathrm{Z}\ \mathrm{Z}\,, \quad \mathrm{X}\ \mathrm{X}\ \mathrm{X}\ \mathrm{X}\ \mathrm{X}\ \mathrm{X}\ \mathrm{X} \tag{1.2}$$

which are interchanged by the transversal Hadamard. Using an identity

$$H = TZT^\dagger = e^{-i\pi Y/8}\ Z\ e^{i\pi Y/8} \tag{1.3}$$

we see that the logical control-Hadamard is possible by replacing the middle Z by the control-Z. The T gate can be noisy as they act on the physical qubits of the Steane code.

This way, we have built a Hadamard measurement routine that is fault-tolerant. Then, a magic state distillation protocol is as follows:

1. Prepare a noisy magic state in the "data" register, and $|0\rangle$ in 6 check registers, and embed them into the Steane code
2. Prepare an ancilla in $|+\rangle$ and implement control-$H^{\otimes 7}$ using Eq. (1.3), where the control is the ancilla and the targets are the physical qubits of the Steane code.
3. Inverse the embedding of the Steane code.
4. Measure the ancilla in the X basis, the check qubits in the Z basis.
5. Upon $+1$ outcome in all 7 measurements, a distilled magic state is in the data qubit.

Let us examine the pattern of errors that may go undetected. There are two possibilities.

- The initial magic state is faulty, and this is undetected due to malfunction of the control-Hadamard.
- The noisy $\pi/4$ rotations induce a logical error.

The first possibility is because a pair of simultaneous errors sandwiching the control-Z can alter the ancilla measurement:

$$\begin{aligned} &|0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes e^{-i\pi Y/8} Y\ Z\ Y e^{i\pi Y/8} \\ &= (Z \otimes I)\left(|0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes e^{-i\pi Y/8}\ Z\ e^{i\pi Y/8}\right). \end{aligned} \tag{1.4}$$

Thus, the first possibility occurs with probability ϵ^3 to leading order. One can easily see that this is the only possibility for weight 2 errors from the control-Hadamard to escape. The second possibility occurs at order ϵ^3 since Steane's code has distance 3. Overall, the protocol operates on 8 qubits (assuming T gates are applied in place), consuming 14 T-gates and 1 state of error rate ϵ, producing 1 output T-state whose error rate is $O(\epsilon^3)$.

It is useful to think of the above protocol as a Hadamard measurement (H-measurement) routine that introduces a new error of order ϵ^3 to the target, and another error of order ϵ^2 to the control. The error on the control is easy to fix; repeat the measurement.[1] The error on the target is inherent to the choice of the inner quantum code, and should thus be overcome by another quantum code.

1.1.2 $[[17, 1, 5]]$ inner code

There exists a distance 5 code on 17 qubits with $H^{\otimes 17}$ being the logical Hadamard. It is an instance of the color code [21, 31]. We include the binary matrix for this code in Appendix A. In a similar way as above, this H-measurement routine has error rate $O(\epsilon^5)$ on the target, and $O(\epsilon^2)$ on the control. By repeating the H-measurement twice using this inner code, the control's error rate becomes $O(\epsilon^4)$. The control error goes undetected only if the initial magic state is faulty. Overall, only weight 5 errors may be undetected. This protocol consumes $17 \times 4 + 1 = 69$ noisy T's.

In fact, we can *pipeline* the two H-measurement routines: First, H-measure a noisy magic state using $[[7, 1, 3]]$ code, and then H-measure the outcome using $[[17, 1, 5]]$ code.[2] Hence, we have obtained a distillation routine with fifth order error suppression that operates on 18 qubits in total, consuming $1+7\times 2+17\times 2 = 49$ noisy T's. It is worth comparing this with the $69 \to 1$ protocol in the preceding paragraph. By using codes of smaller distance in the early stage of the protocol, we obtain a more efficient protocol. We loosely call this modification as a pipelined protocol. The circuit is in Fig. B.2 in Appendix B.

1.2 Repetition outer code

Imagine we have n_{outer} noisy magic states that are to be distilled. Under the stochastic error model, we can think of the noisy magic states as an probabilistic ensemble of n_{outer}-bit strings where 0 denotes a good magic state, and 1 denotes a bad one. The protocol in the previous subsection examines one qubit at a time, and in terms of the bit strings, this amounts to checking individual bits. If our goal is to suppress the overall error to d-th order where $d < n_{\text{outer}}$, the bit-wise check might be overkill. A better way is to devise a measurement routine that can check the parity of several bits.

1.2.1 $[[4, 2, 2]]$ inner code

The simplest case is when $n_{\text{outer}} = 2$ and the desired error suppression is quadratic. If we can measure $H^{\otimes 2}$, then by postselecting on $+1$ outcome the noisy state is projected to the even parity subspace, which is $O(\epsilon^2)$ away from the pair of perfect magic states. We can describe the situation by saying that we have a repetition code on $n_{\text{outer}} = 2$ bits with one parity check. This is an outer code.

[1] This corresponds to having redundant checks in the outer code.

[2] Interestingly, the protocol using Steane code, and the pipelined protocol appear to have deep relation with triply even codes. See Appendix C.

A corresponding inner code should implement control-$H^{\otimes 2}$ to accuracy $O(\epsilon^2)$, both in the target and the control. Meier, Eastin, and Knill have designed such a measurement routine [11]. The four qubit code $[[4,2,2]]$ whose stabilizers are $X^{\otimes 4}$ and $Z^{\otimes 4}$ admits the transversal Hadamard $\bar{H} = H^{\otimes 4}$ as a logical operator. If we choose the logical operators as

$$\begin{cases} \bar{X}_1 = X & X & I & I \\ \bar{Z}_1 = I & Z & Z & I \end{cases} \qquad \begin{cases} \bar{X}_2 = Z & Z & I & I \\ \bar{Z}_2 = I & X & X & I \end{cases} \tag{1.5}$$

then the transversal Hadamard swaps the two logical qubits. Using Eq. (1.3), this means that we can implement control-Swap to accuracy $O(\epsilon^2)$. Now, a trick is to use the control-Swap twice sandwiching the Hadamard:

$$[^C\mathrm{Swap}_{12}]H_1[^C\mathrm{Swap}_{12}] = [^C(H_1 \otimes H_2)]H_1 \tag{1.6}$$

where the superscript C denotes the control that is common for both control-Swaps, and the subscripts 1 and 2 denote the qubits the operator acts on. The extra H_1 does no harm since the magic state is its eigenstate. The obtained control-$H^{\otimes 2}$ is accurate up to error $O(\epsilon^2)$ on the target since the distance of the four-qubit code is 2, and also $O(\epsilon^2)$ on the control due to Eq. (1.4). This is a quadratic distillation protocol operating on 5 qubits, consuming 18 noisy T's to produce 2 outputs.[3]

1.2.2 $[[16,6,4]]$ inner code and pipelining

The classical Hadamard code $[16,5,8]$ has a property that every code word has even overlap with any other code word. By the CSS construction, using these classical codewords as stabilizers, we obtain a $[[16,6,4]]$ code; see Appendix A.1.2 for the stabilizers. We will later show that there is a choice of logical operators such that the transversal Hadamard $H^{\otimes 16}$ implements simultaneous pairwise swaps on the three pairs of logical qubits. This implies that we can measure any even product $\tilde{H}^{\otimes 2}$, $\tilde{H}^{\otimes 4}$, or $\tilde{H}^{\otimes 6}$ of Hadamards on $k_{\mathrm{inner}} = 6$ magic states. For example, we can generalize Eq. (1.6) to

$$[^C\mathrm{Swap}_{12}][^C\mathrm{Swap}_{34}]H_1H_3[^C\mathrm{Swap}_{12}][^C\mathrm{Swap}_{34}] = [^C(H_1 \otimes H_2 \otimes H_3 \otimes H_4)]H_1H_3. \tag{1.7}$$

The H-measurement routine puts quadratic error to the control and quartic error to the target.

Imagine that $n_{\mathrm{outer}} = 6$ magic states are laid on a ring. We measure $H^{\otimes 2}$ on every nearest neighboring pair of the magic states. There are six checks in total. The measurement pattern follows the parity checks of the classical repetition code; there is a redundancy in the checks, which turns out to be necessary. Let us see how this achieves quartic error suppression. In order for an error on one of n_{outer} magic states to pass the measurement routines, the two checks that involve that input state must both be faulty. This process gives an $O(\epsilon^5)$ error, i.e., the probability of both checks being faulty is $O(\epsilon^4)$, so including the error on the input magic state the error is $O(\epsilon^5)$. Note that if we did not have a redundancy in the checks of the outer code, using only 5 checks, one qubit would be checked only once and we would achieve only third order error suppression. More generally, any process involving one or more input magic state errors gives an error which is at most $O(\epsilon^5)$. The dominant error after all the H-measurements is then from the logical error by the H-measurement routine, which happens

[3] Meier, Eastin, and Knill [11] compresses the circuit for the control-$H^{\otimes 2}$ to reduce the number of T's of the protocol from 18 to 10, but we do not explain it here. See also Campbell and O'Gorman [28].

with probability $O(\epsilon^4)$. Overall, the protocol consumes $6 + 6 \times (16 \times 4) = 390$ T's to produce 6 outputs.

We can pipeline the $[[4, 2, 2]]$ code routine in front of the $[[16, 6, 4]]$ code routine to lower the complexity of the distillation circuit. For instance, we can run the three H-measurement routines by the $[[4, 2, 2]]$ code on pairs of magic states (12), (34), and (56), and then run the three H-measurement routines by $[[16, 6, 4]]$ code on pairs of magic states (23), (45), and (61). The number of T's consumed is now $6 + 3 \times (4 \times 4) + 3 \times (16 \times 4) = 246$, while the number of outputs is still 6. It is left to the readers to show that the modified version also achieves quartic error suppression. We have simulated the modified version, and the results can be found in Section 5. The circuit is in Fig. B.1 in Appendix B.

2 Third level of Clifford hierarchy

The protocol above can be straightforwardly generalized to distilling other magic states to implement gates at the third level of the Clifford hierarchy. Consider a state $|\psi\rangle$ on q qubits such that $|\psi\rangle = U|\phi\rangle$ where U is a gate at the third level of the Clifford hierarchy [15], and $|\phi\rangle$ is a stabilizer state. Here we show that any such state $|\psi\rangle$ can be distilled. An example of such a state $|\psi\rangle$ is the magic state to produce a CCZ gate, which is equivalent to the Toffoli gate up to a Hadamard on the target [1].

$$|\psi\rangle = \underbrace{{}^{CC}Z}_{U} \underbrace{\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes 3}}_{|\phi\rangle}. \tag{2.1}$$

As $|\phi\rangle$ is a stabilizer state, we can identify q operators, $S(1), S(2), \ldots, S(q)$, which are products of Paulis, generating the stabilizer group of $|\phi\rangle$, so that $|\phi\rangle$ is the unique (up to global phase) $+1$ eigenstate of those operators. For the CCZ, we see $S(1) = X_1$, $S(2) = X_2$, and $S(3) = X_3$. Then, the state $|\psi\rangle$ is the unique $+1$ eigenstate of the operators

$$W(a) \equiv US(a)U^\dagger, \text{ for } a = 1, \ldots, q. \tag{2.2}$$

These operators $W(a)$ commute with each other by construction, and belong to the second level of the hierarchy, the Clifford group. For the CCZ, we see $W(1) = X_1({}^{C}Z)_{23}$, $W(2) = X_2({}^{C}Z)_{13}$, and $W(3) = X_3({}^{C}Z)_{12}$.

Here is an example protocol for CCZ state distillation using three copies of the $[[4, 2, 2]]$ code, comprising a $[[12, 6, 2]]$ code. We regard the three copies as a single $[[12, 6, 2]]$ code and index the logical qubits by $1, \ldots, 6$. We encode one CCZ state stabilized by $W(a)$ into logical qubits $1, 3, 5$ and another stabilized by $W'(a)$ into logical qubits $2, 4, 6$, where $a = 1, 2, 3$. Consider a variant of Eq. (1.6)

$$\begin{aligned} &[{}^{C}(\mathrm{Swap}_{12}\mathrm{Swap}_{34}\mathrm{Swap}_{56})](W(a)_{135} \otimes I_{246})[{}^{C}(\mathrm{Swap}_{12}\mathrm{Swap}_{34}\mathrm{Swap}_{56})] \\ &= [{}^{C}(W(a)_{135} \otimes W(a)_{246})](W(a)_{135} \otimes I) \end{aligned} \tag{2.3}$$

where the control qubit is common for every gate. The simultaneous control-Swaps are implemented by the control-$H^{\otimes 12}$ on the $[[12, 6, 2]]$ code, where the control-$H^{\otimes 12}$ in turn is implemented by noisy T gates. Thus, we obtain a measurement routine for $W(a) \otimes W'(a)$.[4]

[4] Since $W(1), W(2), W(3)$ are the same up to permutations of qubits, the measurement routine can in fact measure any product $W(a) \otimes W'(b)$ for $a, b = 1, 2, 3$ on the pair of CCZ states.

Then the protocol is to measure $W(1)W'(1)$, $W(2)W'(2)$, and $W(3)W'(3)$. Overall, this protocol takes 2 noisy CCZ states and $3 \times (12 \times 4) = 144$ noisy T gates to produce 2 CCZ states at a lower error rate. Note that this explicit example protocol performs worse than existing ones in terms of non-Clifford gate count [32–34].

By applying the Clifford stabilizers W uniformly at random to a noisy magic state for CCZ, it becomes a mixture of eigenstates of W's. Hence we may assume an error model where an error flips at least one of $W(1), W(2), W(3)$ with probability ϵ. Since the measurement routine puts measurement error at rate $O(\epsilon^2)$ and logical error at rate $O(\epsilon^2)$, the protocol achieves quadratic error reduction for CCZ state. For higher order reduction, one should use inner and outer code of higher distances.

A related discussion on the error model for the T state is given in Section 5.1.

In passing, we note that the Clifford unitary $V_{123} = {}^{(C_1 Z_2)}({}^{(C_1 X_2)} X_2 X_3$ on three qubits $1, 2, 3$ is a stabilizer of the CCZ state $|\mathrm{CCZ}\rangle = \frac{1}{\sqrt{8}} \sum_{a,b,c=0,1} (-1)^{abc} |abc\rangle$. Since the CCZ state is permutation invariant, we obtain six such stabilizers. They do not commute, but any triple of them uniquely determines the CCZ state. The controlled version can be implemented with only four T gates [32] (See also [33]): ${}^{C_0}V_{123} = T_2({}^{C_1}Z_2)T_2({}^{C_0}Z_2)T_2^\dagger({}^{C_1}Z_2)T_2^\dagger({}^{C_0}X_3)$. It might be possible to use these stabilizers with normal codes such as $[[7, 1, 3]]$ and $[[17, 1, 5]]$, but because they do not commute the resulting measurement routine rejects faulty inputs with a probability less than 1, even in the limit $\epsilon \to 0$.

3 Inner Codes

In this section, we find a general class of inner codes that can be used in distillation protocols. On the first read, a reader may wish to skip the discussion on symmetric forms, noting only the magic basis in Definition 3.3, and the construction of codes in Theorem 3.5.

3.1 Symmetric forms over $\mathbb{F}_2$

We consider finite dimensional vector spaces over the binary field $\mathbb{F}_2$. The space $\mathbb{F}_2^n$ is equipped with a symmetric dot product $v \cdot w = \sum_i v_i w_i \in \mathbb{F}_2$. This dot product on $\mathbb{F}_2^n$ is *non-degenerate*, i.e., for any nonzero vector $v \in \mathbb{F}_2^n$ there is a vector $w \in \mathbb{F}_2^n$ such that $v \cdot w \neq 0$. Let $\mathcal{S}$ be a null (self-orthogonal) subspace of $\mathbb{F}_2^n$, on which the dot product identically vanishes. Since $\mathcal{S}$ is null, the dot product of $\mathbb{F}_2^n$ canonically induces a dot product on the quotient space $\mathbb{F}_2^n/\mathcal{S}$ by $[v] \cdot [w] := v \cdot w$ where $[v]$ and $[w]$ denote the equivalence classes (members of the quotient space) represented by v and w, respectively. Let $\mathcal{S}^\perp$ denote the orthogonal complement of $\mathcal{S}$ with respect to the dot product.

Lemma 3.1. *The induced dot product on $\mathcal{S}^\perp/\mathcal{S}$ is non-degenerate.*

Proof. First, we claim that $(\mathcal{S}^\perp)^\perp = \mathcal{S}$. It is clear by definition that $\mathcal{S} \subseteq (\mathcal{S}^\perp)^\perp$. Interpreting the orthogonal complement as the solution space of a system of linear equations, we see that the claim holds by dimension counting. For $[v] \in \mathcal{S}^\perp/\mathcal{S}$, if $[v] \cdot [w] = 0$ for any w, then v belongs to $(\mathcal{S}^\perp)^\perp = \mathcal{S}$, implying that $[v] = 0 \in \mathcal{S}^\perp/\mathcal{S}$. □

For any basis $\{[v^{(1)}], \ldots, [v^{(k)}]\}$ of $\mathcal{S}^\perp/\mathcal{S}$, we consider the symmetric matrix Λ representing the dot product:

$$\Lambda^{ab} = v^{(a)} \cdot v^{(b)}. \tag{3.1}$$

Lemma 3.1 is equivalent to saying that the matrix Λ is non-singular. Any basis change of $\mathcal{S}^\perp/\mathcal{S}$ induces a congruent transformation $\Lambda \to M^T \Lambda M$ where M is the invertible matrix of the basis change. We consider equivalence classes of Λ under the congruent transformations.

Lemma 3.2 (Classification of symmetric forms over $\mathbb{F}_2$)**.** *A non-degenerate symmetric form over $\mathbb{F}_2$ is equivalent to one of the two non-equivalent choices:*[5]

$$I_n = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}, \quad \lambda_n = I_{n/2} \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.2}$$

Proof. The two options are not equivalent since λ_n means that the every vector is self-orthogonal, whereas I_n implies that not every vector is self-orthogonal. For completeness, we give an elementary algorithmic proof by manipulating symmetric matrices.

First, we claim that any symmetric matrix can be brought to a direct sum of I_p and λ_q for some $p \geq 0$ and $q \geq 0$, where q is even. If there is a nonzero diagonal element one can bring this to the top-left by permutation. Gaussian elimination on the first column and row reveals that I_1 is a direct summand. Induction gives a direct summand I_p, and we are left with a symmetric matrix Λ' with the zero diagonal. Any column cannot be zero since Λ' is non-singular, and thus some permutation brings 1 to $(2,1)$ and $(1,2)$ entries of Λ'. Gaussian elimination on the first and second columns and rows reveals a direct summand λ_2. By induction, our first claim is proved.

The second claim is that $I_{p+2} \oplus \lambda_{q-2} \cong I_p \oplus \lambda_q$ whenever $p, q > 0$, whose proof is immediate:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{3.3}$$

Therefore, whenever $p > 0$, we have $I_p \oplus \lambda_q \cong I_{p+q}$. If $p = 0$, there is nothing more to prove. □

The classification motivates the following notion of bases.

Definition 3.3. Given a null subspace $\mathcal{S} \subseteq \mathbb{F}_2^n$, a basis of $\mathcal{S}^\perp/\mathcal{S}$ is called (p,q)**-magic** if the symmetric matrix Λ representing the dot product on $\mathcal{S}^\perp/\mathcal{S}$ among the basis vectors is equal to $I_p \oplus \lambda_q$ for some $p \geq 0$ and $q \geq 0$. We call a magic basis **normal** if $q = 0$, or **hyperbolic** if $p = 0$.

We summarize the results of this section into a theorem.

Theorem 3.4. *For any self-orthogonal subspace $\mathcal{S} \subseteq \mathbb{F}_2^n$, there exists a (p,q)-magic basis for $\mathcal{S}^\perp/\mathcal{S}$, where $p + q = \dim_{\mathbb{F}_2} \mathcal{S}^\perp/\mathcal{S}$. If $p > 0$ and $q > 0$, then a $(p+2, q-2)$-magic basis exists.*

[5] Any finite dimensional symmetric forms over the binary field is the intersection form of a closed topological surface X, defined by the cup product $H^1(X;\mathbb{F}_2) \times H^1(X;\mathbb{F}_2) \to H^2(X;\mathbb{F}_2) \cong \mathbb{F}_2$ of cohomology. Indeed, I_n corresponds to the connected sum of n copies of $\mathbb{R}P^2$, and λ_n the connected sum of $n/2$ copies of 2-torus. The classification in the lemma is the orientability of a given surface, and Eq. (3.3) is expressing the fact that two $\mathbb{R}P^2$'s can be turned into a torus in the presence of another $\mathbb{R}P^2$. We thank Michael Freedman for pointing this out.

3.2 CSS codes from self-orthogonal matrices

It is standard to associate a bit string $v = (v_1, \ldots, v_n)$ to a Pauli operator: $X(v) = X_1^{v_1} \cdots X_n^{v_n}$ where X_j is the Pauli σ^x on qubit j, and $Z(v) = Z_1^{v_1} \cdots Z_n^{v_n}$ where Z_j is the Pauli σ^z on qubit j. The commutation relation is that

$$X(v)Z(w) = (-1)^{v \cdot w} Z(w)X(v). \tag{3.4}$$

The CSS construction of quantum codes applies to a self-orthogonal (null) subspace $\mathcal{S} \subseteq \mathbb{F}_2^n$: For every vector $v \in \mathcal{S}$, we define an X-stabilizer $X(v)$, as well as Z-stabilizer $Z(v)$. The set of equivalence classes of X-type (Z-type) logical operators is then in one-to-one correspondence with $\mathcal{S}^\perp/\mathcal{S}$. The number of logical qubits is thus

$$k = \dim_{\mathbb{F}_2} \mathcal{S}^\perp/\mathcal{S} = n - 2\dim_{\mathbb{F}_2} \mathcal{S}. \tag{3.5}$$

We encode logical qubits by choosing a complete set of logical operators $\tilde{X}^{(j)}$ and $\tilde{Z}^{(j)}$ as follows. Choose a (p, q)-magic basis $\{v^{(1)}, \ldots, v^{(p)}, w^{(1)}, \ldots, w^{(q)}\}$ of $\mathcal{S}^\perp/\mathcal{S}$.[6] Then, we define

$$\begin{cases} \tilde{X}^{(i)} &= X(v^{(i)}) \\ \tilde{Z}^{(i)} &= Z(v^{(i)}) \end{cases} \qquad \text{for } i = 1, \ldots, p,$$

$$\begin{cases} \tilde{X}^{(p+2j-1)} &= X(w^{(2j-1)}) \\ \tilde{Z}^{(p+2j-1)} &= Z(w^{(2j)}) \end{cases} \quad \begin{cases} \tilde{X}^{(p+2j)} &= Z(w^{(2j-1)}) \\ \tilde{Z}^{(p+2j)} &= X(w^{(2j)}) \end{cases} \qquad \text{for } j = 1, \ldots, q/2. \tag{3.6}$$

By definition of the magic basis, these logical operators obey the canonical commutation relation of Pauli operators on k qubits:

$$\tilde{X}^{(a)}\tilde{Z}^{(b)} = (-1)^{\delta_{ab}} \tilde{Z}^{(b)}\tilde{X}^{(a)}. \tag{3.7}$$

Note that the commutation relation can be realized with arbitrary signs $\pm$ in the choice of the logical operators, but induced Clifford logical operators will depend on the signs. We enforce (3.6) in order for the transversal Hadamard $\bar{H} = H^{n_{\text{inner}}}$ to be the logical Hadamard $\prod_{a=1}^{k_{\text{inner}}} \tilde{H}^{(a)}$.

We have defined CSS codes based on self-orthogonal subspaces over $\mathbb{F}_2$:

Theorem 3.5. *Let $\mathcal{S} \subseteq \mathbb{F}_2^n$ be a self-orthogonal subspace with a (p, q)-magic basis of $\mathcal{S}^\perp/\mathcal{S}$. Then, there exists a CSS code on n qubits with $p + q$ logical qubits and a choice of logical operators such that transversal Hadamard $H^{\otimes n}$ implements the logical Hadamards for the logical qubits $1, \ldots, p$, and simultaneously the swaps between the logical qubit $p + 2j - 1$ and $p + 2j$ where $j = 1, \ldots, q/2$.*

We will call a weakly self-dual CSS code normal if a normal magic basis exists, and hyperbolic otherwise. It is possible for a normal code to have an even number of logical qubits, an even number of physical qubits, and an even distance. Every hyperbolic code, however, must have an even number of logical qubits, an even number of physical qubits, and an even distance. For instance, in Sec. 1, the Steane code $[[7, 1, 3]]$ and the $[[17, 1, 5]]$ color code are normal. The $[[4, 2, 2]]$ code and the $[[16, 6, 4]]$ are hyperbolic. We have used $(0, 2)$-magic basis

[6] Here we have abused notation to denote an equivalence class (a member of the quotient space) by a representative.

for $[[4,2,2]]$, and $(0,6)$-magic basis for $[[16,6,4]]$. The "H-code" by Jones [13] is a normal code with parameters $[[k+4,k,2]]$ where k is even. Below, we will mostly use normal codes with $(p,0)$-magic basis for distillation protocols.

We note that the CSS codes derived from a self-orthogonal matrices are not too restrictive. By representing each qubit in any stabilizer code of parameters $[[n,k,d]]$ by Majorana modes, we obtain a weakly self-dual CSS code of parameters $[[4n,2k,2d]]$ [35]. We will briefly review this mapping in Section 4.2, where we will also present other families of such codes with improved rate.

4 Coding Theory and Asymptotic Performance

4.1 Asymptotic Performance

In this section we consider the asymptotic properties of the class of protocols defined above, for appropriate choice of inner and outer codes. We ignore all possibilities of pipelining, and use only a single inner and outer code to define each protocol; this will reduce the question of asymptotic properties to the question of the existence of code families with certain properties.

"Asymptotic" will refer to one of two limits. In the first limit, we consider a family of protocols parametrized by d, the order of reduction in error. An instance in the family reduces error probability from ϵ to a constant times ϵ^d in the limit of small ϵ. We prove that

Theorem 4.1. *There is a family of protocols parametrized by an integer $d \geq 1$ to obtain a d-th order reduction in error, using a total of $\Theta(d)$ physical qubits, producing $n_{\text{outer}} = \Theta(d)$ magic states. The total number of T gates used is $n_T = \Theta(d^2)$, so that the number of T gates per magic state is $\Theta(d)$. The T-gate depth of the circuit is also $\Theta(d)$, where the T-gate depth refers to the circuit depth assuming that an arbitrary Clifford can be executed in depth 1.*

In the second limit, we fix d and consider a family of protocols parametrized by n_{outer}, the number of magic states produced. We prove that

Theorem 4.2. *For any odd $d \geq 5$, there is a family of protocols using $n_{\text{outer}} \cdot (1 + o(1))$ physical qubits, producing n_{outer} magic states with a d-th order reduction in error. The total number of T gates used is*

$$n_T = d(1 + o(1))n_{\text{outer}}. \tag{4.1}$$

The reason d is odd is that the minimal weight of an error that is not caught in the protocol due to wrong H-measurement outcomes is always odd.

Given one particular protocol with $\epsilon_{\text{out}} = C\epsilon_{\text{in}}^d$ consuming n_T/n_{outer} T gates per output, an infinite family of protocols can be defined by concatenation with itself. For this concatenated family, the number of T gates to achieve an arbitrarily small error rate δ in output magic states scales like $O((\log 1/\delta)^\gamma)$ where the scaling exponent [10, 12] is

$$\gamma = \log_d(n_T/n_{\text{outer}}). \tag{4.2}$$

Smaller values of γ reflect asymptotically more efficient distillation protocols. The triorthogonal codes [12] achieve $\gamma \to \log_2(3)$, and "multilevel" protocol [13] achieves $\gamma \to 1^+$. We will comment on this multilevel protocol in Discussion 6. It was conjectured that no protocol could achieve $\gamma < 1$ [12]. Both families in Theorems 4.1 and 4.2 achieve $\gamma \to 1^+$.

We note that the measure γ slightly underestimates the T-count efficiency of the family in Theorem 4.1. In order to achieve an arbitrary small final error rate δ from a fixed initial error rate, say, $\epsilon = 0.01$, we can pick a member P_d of the family of error reduction degree d such that $\delta > C_d(\epsilon/2d^2)^d$. Here C_d is the leading coefficient of the output error probability of the protocol P_d, which is at most the number of ways that weight d errors occur among $n_T = O(d^2)$ T gates; $C_d \leq \alpha d^{2d}$ for some $\alpha > 0$ independent of d. For the condition $\delta > C_d(\epsilon/d^2)^d$, it suffices that $d > (\log(1/\delta) + \log\alpha)/\log(1/\epsilon)$. We initially distill magic states to suppress the error rate from ϵ to $\epsilon' = \epsilon/d^2$, by using a concatenated protocol P_{init}. This takes $n_{\text{init}} = O(\log d)^\gamma$ input magic states per output magic states for some $\gamma > 1$. We can then feed P_d with the outputs from P_{init} at error rate ϵ'. It follows that

$$n_T/n_{\text{outer}} = O(d) \cdot n_{\text{init}} = O(\log(1/\delta)(\log\log 1/\delta)^\gamma) \tag{4.3}$$

magic states at error rate ϵ suffice to achieve final accuracy δ. Thus, the scaling of n_T/n_{outer} is linear in $\log(1/\delta)$ up to a logarithmic correction. (One can iterate the argument recursively to further slow down the dependency on $1/\delta$.)

Theorem 4.2 will use normal codes. The reduced number of T gates required to implement checks with a normal code is essential to obtaining the number of T gates in the theorem (we would need roughly twice as many using hyperbolic codes). This explains why d is chosen odd. The case of $d = 1$ is of course trivial: no codes are needed. Thus, the reader may wonder why the case $d = 3$ is not used; this is explained further below.

Theorems 4.1, 4.2 will follow almost immediately given certain families of inner and outer codes obeying certain properties of the codes that we define below. We will prove these theorems given these properties in this subsection and we construct families of inner and outer codes with these properties in subsections 4.2, 4.3.

Consider first the inner code. This code will have k_{inner} logical qubits and n_{inner} physical qubits. The distance of the inner code will be at least d. Consider then the effect of errors in the T gates inside the inner code; i.e., in the T gates acting on the encoded state. To obtain d-th order reduction in error, it suffices to consider the case that fewer than d errors occur in such T gates. Since the inner code distance is at least d, these errors cannot produce a logical error. There is one way, however, in which these errors can have an effect without being detected by the inner code. It is possible that a pair of errors act inside the inner code, both on T gates acting on the same qubit. The effect of these errors is to cause an error in the check being measured by the inner code, i.e., if the check was measuring a given product of W operators specified by the outer code, we instead measure the opposite sign; we call this a "measurement error".

The possibility of measurement errors affects some of the properties that we require of the outer code. Of course we need the outer code to have distance at least d, as otherwise a pattern of fewer than d errors in the input magic states could cause an undetectable error, but this is not sufficient. It is necessary that a pattern of fewer than d errors causes enough checks to be violated so that even a small number of measurement errors will lead to an error detected by the code. This is defined by the property of "sensitivity" that we now define.

The outer code will have m parity checks, encoded in an m-by-n_{outer} parity check matrix M, where each row of the matrix indicates a given check. We can measure rows of this matrix with even weight using an hyperbolic inner code and rows with odd weight using a normal inner code. For simplicity we will either have all rows have even weight or have all rows use

odd weight so that we can use the same inner code for all checks. (More generally, one could use both a hyperbolic and a normal code.) Then, this inner code must have k_{inner} greater than or equal to the maximum row weight of M. The difference between the row weight of M and k_{inner} must be an even number. In this case, we say that the inner code can implement the checks of the outer code.

Definition 4.3. An m-by-n_{outer} parity check matrix M for a classical linear code is said to be $(\tilde{d}, s)$**-sensitive** if any nonzero bit vector v of length n_{outer} with $|v| \leq \tilde{d}$, we have $|Mv| \geq s$. That is, for any such vector, the number of violated parity checks is at least s.

We emphasize that sensitivity is a property of the check matrix of the outer code, rather than the codewords of the outer code, and in some examples the rows of the check matrix may be linearly dependent. A $(\tilde{d}, s)$-sensitive parity check matrix is $(\tilde{d}-1, s)$-sensitive by definition.

Lemma 4.4. *Given an m-by-n_{outer} parity check matrix M such that $2|Mv| + |v| \geq d$ for any nonzero v (e.g. $(d-1, \frac{d-1}{2})$-sensitive M), and given an inner code of parameters $[[n_{\text{inner}}, k_{\text{inner}}, d]]$ that can implement the checks defined by M, the protocol yields d-th order reduction in error. The protocol overall takes n_{outer} noisy magic states, $2n_{\text{inner}}m$ noisy T gates when the inner code is normal, or $4n_{\text{inner}}m$ when hyperbolic, and outputs n_{outer} magic states.*

Proof. Any error pattern inside the inner codes with weight less than d cannot cause a logical error. Thus, if an error pattern inside the inner code does not violate a stabilizer of the inner code, it either has no effect or it leads to an error in measurement of a check of the outer code; the latter possibility requires at least two errors inside the inner code. Any input state with $|v| \geq 1$ errors will violate at least $|Mv| \geq (d - |v|)/2$ checks of the outer code. If no violation of these checks is detected, there must be at least $2|Mv|$ errors on T gates inside the inner code. Thus, there must be at least d errors in total.

The input T gate and state count is clear from Section 1. □

We now define some asymptotic properties of the codes needed.

Definition 4.5. A family of quantum error correcting codes with increasing number of qubits n has **good rate** if the number of encoded qubits k is $\Theta(n)$ and has good distance if the distance d is $\Theta(n)$.

Definition 4.6. Given a family of outer codes with increasing n_{outer}, we say that this family has **good sensitivity** if each code in the family is $(\tilde{d}, s)$-sensitive for $\tilde{d} = \Theta(n_{\text{outer}})$ and $s = \Theta(n_{\text{outer}})$.

Definition 4.7. Given a family of outer codes with increasing n_{outer}, we say that this family has **good check rate** if the parity check matrix is m-by-n_{outer} with $m = \Theta(n_{\text{outer}})$.

Proof of Theorem 4.1. In subsection 4.2, we show that families of both hyperbolic and normal inner codes with good rate and distance exist and in subsection 4.3 we show that families of outer codes with good check rate, good sensitivity, and even row weight exist. Combining these results with Lemma 4.4, Theorem 4.1 follows. □

Proof of Theorem 4.2. In subsection 4.2, we show that, for any d, there exist families of both hyperbolic and normal inner codes with increasing n_{inner} such that $k_{\text{inner}}/n_{\text{inner}} \to 1$. To prove this theorem, we will only need the result for normal inner codes. Consider some code from this family with given $k_{\text{inner}}, n_{\text{inner}}$. In subsection 4.3 we show Lemma 4.15 which we reproduce here:

Lemma. Given integers $\tilde{d}, w \geq 1$ and $s \geq 2$, there exists an $m \times n_{\text{outer}}$ parity check matrix M that is $(\tilde{d}, s)$-sensitive where $m = n_{\text{outer}} \cdot s/w$ and every row of M has weight w exactly.

Choose $w = k_{\text{inner}}$. Choosing $\tilde{d} \geq d-1$ and $s = (d-1)/2$, this gives us an outer code such that the checks can be performed by the given inner code and we need to perform $n_{\text{outer}}(s/w) = n_{\text{outer}}(s/k_{\text{inner}})$ checks with the inner code. Each such check with the inner code requires using $2n_{\text{inner}}$ T gates, so that the total number of T gates needed to perform the checks with the inner code is equal to $2n_{\text{outer}}s(n_{\text{inner}}/k_{\text{inner}})$. Additionally, we need to perform n_{outer} T gates to create the input magic states to the outer code. Thus, the total number of T gates is

$$\begin{aligned} n_T &= n_{\text{outer}}(1 + 2sn_{\text{inner}}/k_{\text{inner}}) \\ &= n_{\text{outer}}(d + (d-1)(n_{\text{inner}}/k_{\text{inner}} - 1)). \end{aligned} \tag{4.4}$$

Taking n_{inner} large so $n_{\text{inner}}/k_{\text{inner}} \to 1$, we conclude $n_T \to n_{\text{outer}}d$. □

We can now see better why we needed $d \geq 5$ in Theorem 4.2. This is because for $d = 3$, we have $s = 1$ and Lemma 4.15 does not apply. The reader will see later why the case $s = 1$ is excluded from that lemma; roughly, this is because in this case, each bit participates in only a single check and we would lack certain expansion properties for a certain graph defined later.

4.2 Inner Codes

In this subsection, we give asymptotic constructions of inner codes. There are at least two constructions of weakly self-dual codes with good rate and distance in the literature. We review these, before giving an alternative probabilistic proof which has some advatanges.

First, in Ref. [36], it is shown that given any ratio d/n, one can find a family of weakly self-dual CSS codes with n qubits and distance d and given ratio d/n achieving a rate $k/n \to 1 - 2H_2(d/n)$, where H_2 is the binary entropy function. The codes found in that paper all are hyperbolic codes. However, we can obtain normal codes from them by a "puncturing procedure" (see also [37, Sec. 3.5]):

Definition 4.8. Given a hyperbolic weakly-self-dual CSS code C on n qubits with k logical qubits, define a "punctured code" C' as follows. Choose a qubit i (the code C' may depend upon the choice of i). Write the stabilizer generators of C such that only one X-type and one Z-type generator is supported on i. Define C' by removing qubit i and removing the stabilizer generators support on i. Then C' has $n' = n - 1$ qubits and $k + 1$ logical qubits. The code C' is a normal code by construction.

If C is non-degenerate with distance d, then C' has distance $d' \geq d - 1$. More generally, $d' + 1$ is greater than or equal to the minimum weight of an operator which commutes with the stabilizer group of C, because given an X-type logical operator O in C' then either O or

OX_i must commute with the stabilizer group of C. Indeed, one may show that puncturing the codes of Ref. [36] reduces the distance by at most 1.

The only disadvantage of this proof is that it is a greedy proof that we not know how to implement efficiently. While it would be desirable to find an explicit family of codes achieving this rate, we do not know how to do this. However, another construction in the literature is a randomized construction which allows us to give codes which, with high probability, have the desired distance. Unfortunately, this construction will only achieve $k/n \to 1/2 - H_2(2d/n)$. This construction uses a general method to construct weakly self-dual CSS codes in Ref. [35].

Consider a stabilizer code C_{qubit} which acts on n physical qubits and has k logical qubits and distance d. From this code, one can derive a code for Majorana fermions $C_{Majorana}$ which acts on $4n$ Majorana modes and has k logical qubits and distance $2d$, where now the distance refers to minimum weight of a product of Majorana operators that is a logical operator. The code $C_{Majorana}$ is derived in the following way: For each physical qubit of C_{qubit}, one introduces four Majorana modes, $\gamma_0, \gamma_1, \gamma_2, \gamma_3$, and declares that the product $\gamma_0\gamma_1\gamma_2\gamma_3$ is a stabilizer of $C_{Majorana}$. For each stabilizer of C_{qubit}, one defines a stabilizer of $C_{Majorana}$ by replacing an operator X on a qubit by $i\gamma_0\gamma_1$, Y by $i\gamma_0\gamma_2$, and Z by $i\gamma_0\gamma_3$. The stabilizer generators of $C_{Majorana}$ are given by bit strings of length $4n$ such that the dot product over $\mathbb{F}_2$ of any pair of such bit strings is 0. Thus, from $C_{Majorana}$, one can define a weakly self-dual CSS code C_{wsd} with $4n$ physical qubits, $2k$ logical qubits and distance $2d$. Since a randomized construction (see, for example, Eq. 7.200 of Ref. [38]) gives stabilizer codes C_{qubit} with $k/n \to 1 - H_2(d/n) - (d/n)\log_2(3)$, mapping these stabilizer codes C_{qubit} to weakly-self dual codes C_{wsd} gives $k/n \to (1/2)[1 - H_2(2d/n) - (2d/n)\log_2(3)]$. Since the randomized construction gives a lower bound to the weight of any operator commuting with the stabilizer group, we can puncture these codes and reduce the distance by at most 1.

Here we give another proof of the existence of such good weakly self dual-codes. This will lead to rate $k/n \to 1 - 2H_2(d/n)$. For any fixed distance d, one can obtain families of stabilizer codes with n physical qubits and k logical qubits with the ratio $k/n \to 1$ as $n \to \infty$. While this improvement is only by constant factors over the construction via Majorana codes, it will lead to nice asymptotic expressions for the number of T-gates, n_T, required to attain d-th order suppression in error. It is also a randomized construction, showing that codes in a certain ensemble have the desired properties with high probability.

Define a random ensemble of c-by-n self-orthogonal matrices as follows, where a matrix M is defined to be self-orthogonal if $MM^T = 0$. Choose the first row of the matrix to be the all 1s vector $\vec{1}$. Choose the second row uniformly at random subject to the constraint that it have vanishing dot product with the first row. Continue in this fashion, choosing the j-th row uniformly at random subject to the constraint that it have vanishing dot product with the first $j-1$ rows. (Remark: the requirement that the first row be the all 1s vector is simply chosen to simplify some notation, so that we do not need to add the requirement that each row have even weight.)

Lemma 4.9. *Consider a fixed n-component vector v, with $v \neq 0$ and $v \neq \vec{1}$. For a random c-by-n self-orthogonal M, the probability that $Mv = 0$ is at most $2^{-c+1} + 2^{-n+c+1}$.*

Proof. Let $w_1, \ldots, w_c$ be the rows of M. Let V_j be the self-orthogonal subspace which is the span of the first j rows of M. We will estimate the desired probability by a union bound, considering separately the event that $v \in V_c^\perp$ and $v \notin V_c$, and the event that $v \in V_c^\perp$ and $v \in V_c$.

Consider the first event. Let $j > 1$. Then

$$\Pr[w_j \cdot v = 0 | v \notin V_{j-1}] = \frac{1}{2}, \tag{4.5}$$

because the constraint that $(v, w_j) = 0$ is independent of the constraints on the vector w_j. Thus, for any k,

$$\Pr[v \in V_k^\perp \text{ and } v \notin V_k] \leq \prod_{j=2}^{k} \frac{1}{2} = 2^{-k+1}. \tag{4.6}$$

For $k = c$, we find in particular that

$$\Pr[v \in V_c^\perp \text{ and } v \notin V_c] \leq 2^{-c+1}. \tag{4.7}$$

Now we estimate the probability of the second event. Note that if $v \in V_c$, there is a least j such that $v \in V_j$. So,

$$\Pr[v \in V_c^\perp \text{ and } v \in V_c] \leq \sum_{j=2}^{c} \Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}]. \tag{4.8}$$

We have

$$\begin{aligned} &\Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \\ &= \Pr[v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \cdot \Pr[v \in V_j \mid v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \\ &\leq 2^{-j+2} \Pr[v \in V_j \mid v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}], \end{aligned} \tag{4.9}$$

where we used Eq. (4.6).

Now we estimate the probability $\Pr[v \in V_j \mid v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}]$. This is possibly nonzero only if $v \cdot v = 0$. Consider the space of all n-component vectors modulo vectors in V_{j-1}; this quotient space has dimension at least $n - (j-1)$. Let π be the natural map from the space of all vectors to this quotient space. The vector πv is nonzero by assumption. The vector w_j is subject to at most $j-1$ independent constraints from V_{j-1}. Consider the space of possible πw_j, given that w_j obeys those constraints; this space has dimension at least $n - 2(j-1)$ and so the probability that a random vector in this space is equal to πv is at most $2^{-(n-2j+2)}$. Hence, $\Pr[v \in V_j \mid v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \leq 2^{-(n-2j+2)}$, so $\Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \leq 2^{-n+j}$. So by Eq. (4.8),

$$\Pr[v \in V_c^\perp \text{ and } v \in V_c] \leq \sum_{j=2}^{c} 2^{-n+j} \leq 2^{-n+c+1}. \tag{4.10}$$

By a union bound, adding probabilities in Eqs. (4.7,4.10), the lemma follows. □

Lemma 4.10. *Let n, c, d be such that*

$$(2^{-n+c+1} + 2^{-c+1}) \sum_{w=1}^{d} \binom{n}{w} < 1. \tag{4.11}$$

Then, there exists a c-by-n matrix M such that $MM^T = 0$ and such that $Mv \neq 0$ for any $v \neq 0$ with v having Hamming weight at most d.

Proof. This follows from Lemma 4.9 and by a first moment bound. For a random M from the above ensemble, the expected number of vectors $v \neq 0$ with Hamming weight at most d such that $Mv = 0$ is at most $(\sum_{w=1}^{d} \binom{n}{w}) \cdot (2^{-n+c+1} + 2^{-c+1})$. □

Lemma 4.11. *For any fixed d, one can find a family of M with increasing n such that the ratio c/n tends asymptotically to zero and such that Eq. (4.11) is obeyed. Hence, for any distance d, one can find a family of hyperbolic or normal weakly self-dual CSS codes such that the ratio $k_{\text{inner}}/n_{\text{inner}} \to 1$ as $n_{\text{inner}} \to \infty$.*

Proof. Immediate for the hyperbolic case. Since the lemma 4.9 upper bounds the probability that an operator commutes with the stabilizer group one can also puncture these codes to obtain a normal code. □

4.3 Outer Codes

In this subsection, we construct families of outer codes with good check rate and sensitivity. We begin with a randomized construction, and then show how to construct explicit families using previous results in coding theory.

Lemma 4.12. *There exist families of outer codes with good check rate and sensitivity and even row weight. Similarly, there exist families of outer codes with good check rate and sensitivity and odd row weight.*

Proof. We only give a proof for check matrices of even row weight. The proof for odd row weight case is completely analogous.

Consider a random m-by-n_{outer} parity check matrix M. Let $\tilde{d} = n_{\text{outer}} - 1$. Choose each row independently but with the constraint that it should be of even weight. For any vector v with $|v| \leq \tilde{d}$, the syndrome vector Mv has independent entries from the uniform distribution. Thus, the probability that $|Mv| \leq s$ for $s \leq m/2$ is bounded by

$$2^{-m} \sum_{i \leq s} \binom{m}{i} = \mathcal{O}^*(2^{m(H(s/m)-1)}),$$

where $H(p) = -p \log_2(p) - (1-p) \log_2(1-p)$ is the binary entropy function, and $\mathcal{O}^*$ hides polynomial factors. The number of such vectors v is bounded by $2^{n_{\text{outer}}}$. By a union bound, the probability that there is an error vector v of weight less than $\tilde{d}$ such that the syndrome has weight less than s is bounded by

$$\mathcal{O}^*(2^{m(H(s/m)-1)}) 2^{n_{\text{outer}}} = \mathcal{O}^*(2^{n_{\text{outer}}(1+(m/n_{\text{outer}})(H(s/m)-1))}).$$

For sufficiently large ratio m/n_{outer} and sufficiently small ratio s/m, this quantity is exponentially small in n_{outer}. □

The above randomized construction is very similar to randomized constructions of classical codes with good rate and distance, where we define

Definition 4.13. A family of classical error correcting codes with increasing number of bits n has **good rate** if the number of encoded bits k is $\Theta(n)$ and has **good distance** if the distance d is $\Theta(n)$.

That is, even though we are considering very different properties (number of violated checks rather than distance of the code), the first moment argument above is very similar to standard first moment arguments to construct such codes with good rate and distance, with some additional technicalities required to ensure even weight of the parity checks. This is not a coincidence. As we now show, given a family of codes with good rate and distance, one can construct a family of codes with good check rate and sensitivity.

Lemma 4.14. *Let C be a classical error correcting code that encodes k bit messages into n bit codewords. Let C have distance d. Let $v_1, \ldots, v_k$ be a basis for the codewords of C. Let M be the n-by-$(k+1)$ matrix whose columns are the vectors $v_1, \ldots, v_k, w$ where $w = v_1 + \ldots + v_k$. Then, all rows of M have even weight and M is a parity check matrix for a code with $n_{\text{outer}} = k+1$ bits which is $(\tilde{d}, s)$ sensitive with $s = d$ and $\tilde{d} = n_{\text{outer}} - 1$. Thus, the code with parity checks encoded by M has only two codewords (the all 0 vector and the all 1 vector) and any message which is not a codeword will violate at least d checks.*

Proof. For any $(k+1)$-bit vector v, the vector Mv is a codeword of C. If v is nonzero and is not equal to the all 1 vector, then Mv is a nonzero codeword of C and hence has weight at least d. □

Since $n_{\text{outer}} = k+1$, in order to obtain an even n_{outer}, if C has k even, we can simply define a new code C' which encodes $k-1$ bit messages into n bit codewords by using any $(k-1)$-dimensional subspace of the codewords of C, in this way obtaining a parity check matrix for a code with $n_{\text{outer}} = k - 1 + 1 = k$.

Using lemma 4.14, we can construct explicit families of codes with good check rate and good sensitivity given any explicit family of codes with good rate and good distance. As an example of such a code family, we can use the expander codes of Ref. [39].

Lemma 4.15. *Given integers $\tilde{d}, w \geq 1$ and $s \geq 2$, there exists an $m \times n_{\text{outer}}$ parity check matrix M that is $(\tilde{d}, s)$-sensitive where $m = n_{\text{outer}} \cdot s/w$ and every row of M has weight w exactly.*

Proof. A parity check matrix M defines a bipartite graph G, often called a Tanner graph. One set of vertices of the graph (which we call B labeled by the columns of M) corresponds to bits of the code and the other set (which we call C labeled by the rows of M) corresponds to checks, with an edge between a pair of vertices if M is nonzero in the corresponding entry. Equivalently, given such a bipartite graph G, this defines a parity check matrix. We claim that given a bipartite graph with all vertices in B having degree s and all vertices in C having degree w and with girth $> 2\tilde{d}$, the corresponding parity check matrix defines a code with the desired properties. Once we have shown this, the lemma follows, since Ref. [40] shows the existence of such graphs.

Note first that the degree of vertices in C corresponds to the row weight of M. Next, note that if all vertices in C have degree w and all in B have degree s, then $m = |C| = n_{\text{outer}} \frac{s}{w}$ with $n_{\text{outer}} = |B|$.

To prove the claim, let $V \subseteq B$ be a nonempty set of erroneous bits. By assumption, $1 \leq |V| \leq \tilde{d}$. Consider a subgraph H of G defined by all vertices of V and its neighbors. By the girth condition on G, the subgraph H has to be a collection of disjoint trees. Thus, it suffices to prove the claim in case where H is connected. If $|V| = 1$, then the error violates s checks, and we are done. If $|V| \geq 2$, let $v_1, v_2 \in V$ be a pair that are the furthest apart. The

choice of the pair ensures that each of v_1 and v_2 has $s-1$ leaves attached to it. Therefore, V violates at least $2s-2 \geq s$ checks. □

Note that the ratio $m/n_{\text{outer}} = s/w$ in lemma 4.15 is the best possible, because each bit must participate in at least s checks (i.e., every column of the parity check matrix must have weight at least s). Note also that though the existence of desired graphs is guaranteed, they might be too large in practice; $w^s \leq n_{\text{outer}} \leq \text{poly}(w^s)$ [40]. However, one does not have to be too strict on the biregularity of the graph in practice. If small violation of the biregularity gives a much smaller graph, then it might be more useful.

5 Numerical Simulation

In this section, we give results of numerical simulations. We begin by explaining the error model we used for simulations. We then explain two protocols that we simulate that are not explained previously; one of these protocols uses a $[[21, 3, 5]]$ code. Then we give the simulation results. One interesting result of the simulation is how little effect the subleading terms have, even at fairly large noise values.

5.1 Magic state fidelity

When we inject a magic state μ for a $\pi/4$ rotation into a quantum circuit, there is a probability for correction K by angle $\pi/2$ to be applied. If we represent the overall procedure by a quantum channel $\mathcal{C}_\mu$, it is $\mathcal{C}_\mu(\rho) = \Pi_+(\rho\otimes\mu)\Pi_+ + K\Pi_-(\rho\otimes\mu)\Pi_- K^\dagger$, where $\Pi_\pm$ denotes the measurement combined with a control-Pauli on the magic state and a target data qubit. Let $|\mu_0\rangle$ be the ideal magic state, and $|\mu_0^\perp\rangle$ be the orthogonal state. Then, it is straightforward to calculate that $\mathcal{C}_{|\mu_0\rangle\langle\mu_0^\perp| + |\mu_0^\perp\rangle\langle\mu_0|}(\rho) = 0$.

This implies that for *any* initial approximate magic state μ, the result of the injection is the same as if μ had been through a twirling channel $\mathcal{E}$ that dephases the magic state in the basis $\{|\mu_0\rangle, |\mu_0^\perp\rangle\}$:

$$\mu = \begin{pmatrix} 1-\epsilon & * \\ * & \epsilon \end{pmatrix} \xrightarrow{\mathcal{E}} \begin{pmatrix} 1-\epsilon & 0 \\ 0 & \epsilon \end{pmatrix}. \tag{5.1}$$

The twirled state is ϵ away from the ideal state in the trace distance,[7] resulting in error at most ϵ to the quantum circuit's outcome. The error ϵ can be expressed by the squared fidelity[8] as

$$1-\epsilon = F^2(\mu_0, \mu) = \langle\mu_0|\,\mu\,|\mu_0\rangle = \langle\mu_0|\,\mathcal{E}(\mu)\,|\mu_0\rangle\,. \tag{5.2}$$

This formula is convenient in that it yields the same answer regardless of whether or not twirling is applied to μ (this is the last equality in the above formula). When a state μ_n that approximates $\mu_0^{\otimes n}$ is injected, the error from this multi-qubit magic state is given by $1 - F^2(\mu_0^{\otimes n}, \mu_n)$. Note that $F^2(\mu_0, \mu)$ is linear in μ. Below, we use $1 - F^2$ as the probability of error to report our simulation results.

[7] The trace distance is defined as $T(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_1$.

[8] The fidelity is defined as $F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$, which is equal to $|\,\langle\rho|\sigma\rangle\,|$ for pure ρ and σ.

5.2 Error Models

The typical model to analyze distillation protocols is the stochastic error model. In typical distillation protocols, one has only a single output magic state, and so one is interested in the probability that the output magic state has an error as a function of the input, conditioned on no error being detected by the code; the error probability is a ratio of polynomials in ϵ, with the leading term being of order ϵ^d for some d, with an integer coefficient.

For our purposes, since the codes used are fairly large, enumeration of all possible error patterns becomes difficult, especially if one wishes to go beyond leading order in ϵ. For this reason, we use numerical simulation. One could simulate a mixed state, using a quantum channel to describe an approximate T-gate; however, this is numerically prohibitive and so we prefer to use an approach that involves only pure states. One could numerically simulate pure states using the stochastic error model by choosing errors to occur with probability p, and sampling the output error probability. However, this simulation also becomes difficult, precisely because the codes lead to a high suppression in the error. For example, if the target error probability is 10^{-10}, one would require $\sim 10^{10}$ samples, with a fairly large number of qubits needed to be simulated in each run, to determine the output error probability accurately.

While there may be ways to overcome this sampling issue using importance sampling, we use another method. Instead of rotating by either $\pi/4$ or by $5\pi/4$ as in the stochastic error model, each T gate rotates by an angle chosen uniformly in the interval $[\pi/4-\theta, \pi/4+\theta]$, for some angle $\theta > 0$. Then, conditioned on the code not detecting an error, we determine the error in the output state.

In fact, the model with input angles $[\pi/4-\theta, \pi/4+\theta]$ and the stochastic error model describe the same average input state, assuming an appropriate choice of ϵ and θ.

$$\mu = \frac{1}{2\theta}\int_{[-\theta,+\theta]} \mathrm{d}x \underbrace{\begin{pmatrix} \cos^2\frac{x}{2} & \sin\frac{x}{2}\cos\frac{x}{2} \\ \sin\frac{x}{2}\cos\frac{x}{2} & \sin^2\frac{x}{2} \end{pmatrix}}_{\rho_x} = (1-\epsilon)\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{\mu_0} + \epsilon \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}}_{\mu_1}$$

$$\frac{1}{2} - \frac{\sin\theta}{2\theta} = \epsilon \tag{5.3}$$

Hence, one wants $\epsilon \approx \theta^2/12$. (We emphasize that this is in a notation where θ is the rotation angle in the Bloch sphere; the T-gate is a rotation by $\pi/4$, not by $\pi/8$.) In the stochastic error model with small ϵ, one must do roughly $1/\epsilon$ runs to obtain meaningful statistics, while here, one needs only a constant number of runs. The reason is as follows. Since θ is small, the simulated circuit can be approximated by an analytic series in θ, and the linear term amounts to a single error, which is projected out by the post-selection on measurement outcomes as our protocol always has $d \geq 2$. Thus, in our post-selected simulation, a circuit with $\rho_{x\in[-\theta,\theta]}$ is equivalent at leading order to a circuit with ρ'_x where ρ'_x is ρ_x with the linear term in x dropped. Then, the distance from n-sample average of ρ'_x to μ is $O(\epsilon/\sqrt{n})$, whereas the distance from n-sample average of μ_i $(i=0,1)$ to μ is $O(\sqrt{\epsilon/n})$. The acceptance probability depends on the fidelity to the ideal magic state μ_0, which is $1-O(\epsilon) = 1-O(\theta^2)$ in any case.

5.3 Other Protocols

5.3.1 $[[16, 2, 4]]$ Inner Code

In subsubsection 1.2.2 we explained a protocol using a $[[16, 6, 4]]$ inner code. This required using a total of 17 physical qubits, namely 16 for the code and one ancilla. We can also modify this inner code to a $[[16, 2, 4]]$ inner code, by turning some of the logical operators into checks. This inner code suffices to implement the H-measurements on pairs of states $(23), (45), (61)$ and so it can implement the checks of the outer code used in subsubsection 1.2.2. Using a $[[16, 2, 4]]$ inner code, if we want to have $n_{\text{outer}} = 6$, we need a total of 21 physical qubits, since we need 16 for the code, plus 4 for the logical qubits not encoded in the code, plus one ancilla. Thus, this requires additional physical qubits compared to the $[[16, 6, 4]]$ code. The reason for considering the $[[16, 2, 4]]$ code in numerics is to see if it reduces the prefactor in the error, since the $[[16, 2, 4]]$ code has fewer logical operators than the $[[16, 6, 4]]$ code. The number of T gates in the protocol using $[[16, 2, 4]]$ is the same as that using $[[16, 6, 4]]$. We pipeline the protocol with the $[[16, 2, 4]]$ inner code in the same way as we did with the $[[16, 6, 4]]$ inner code. See Fig. B.3 in Appendix B.

5.3.2 $[[21, 3, 5]]$ Inner Code

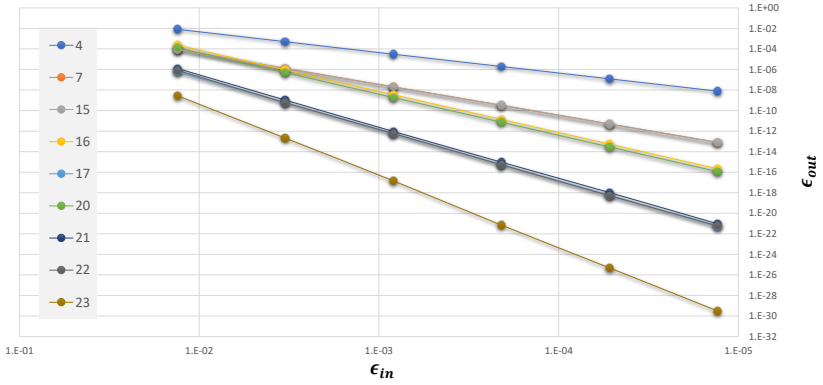
Another inner code that we used is a $[[21, 3, 5]]$ inner code, described in Appendix A. This allows us to obtain fifth order reduction in error. We used $n_{\text{outer}} = 4$ with the outer code having check matrix

$$M = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}. \tag{5.4}$$

It uses 4 checks for 4 qubits. This matrix is not $(4, 2)$-sensitive, but is sufficient to achieve fifth order reduction in error since $2|Mv| + |v| \geq 5$ for every nonzero v. This protocol consumes $4 + 4 \times (21 \times 2) = 172$ T's to produce 4 outputs T states. See Fig. B.4 in Appendix B.

A simple pipelining can reduce the noisy T gate count compared to this protocol. Distill three independent magic states using $[[7, 1, 3]]$ inner code. (The outer code is trivial in this case.) The three distilled magic states is then pipelined into the $[[21, 3, 5]]$ inner code. This produces 3 magic states with error $O(\epsilon^5)$, consuming, per output, 28 T gates and one T state with error ϵ.

Without the pipelining in the preceding paragraph, but using weight 3 checks from $[[21, 3, 5]]$, we can find an outer code that is $(4, 2)$ sensitive using $(2/3)n_{\text{outer}}$ checks. This produces n_{outer} magic states with fifth order error suppression, consuming $42 \cdot 2/3 = 28$ T gates and one T state, per output. The smallest such outer code is explained in the appendix, where it is called the Petersen graph code. For this outer code, $n_{\text{outer}} = 15$, so the overall protocol consumes 435 T's to produce 15 outputs. It is possible to further reduce the number of input T's, by replacing some of the early checks with a smaller code e.g. $[[15, 7, 3]]$ code, which can be obtained by puncturing $[[16, 6, 4]]$ code.



Protocol	d	n_T/n_{outer}	n_{outer}	Prefactor
"4"	2	9	2	45×10^0
"7"	3	15	1	35×10^0
"15"	3	15	1	35×10^0
"16"	4	41	6	7.3×10^3
"17"	5	49	1	1.4×10^3
"20"	4	41	6	3.9×10^3
"21"	5	29	3	2.5×10^3
"22"	5	43	4	1.4×10^3
"23"	7	95	1	49×10^3

Figure 5.1: Results of numerical simulations. ϵ_{in} represents input error; this is the error ϵ of Eq. (5.3) for the given θ. ϵ_{out} is the average of $1 - \langle\mu_0|\,\mu\,|\mu_0\rangle$ over runs. The fitting function is $\epsilon_{out} = C\epsilon_{in}^d$. The graph confirms that the leading term dominates the output error probability in the simulated regime. The numbers labelling curves indicate the number of physical qubits, not including the ancilla qubit. Specifically, "4" indicates protocol using $[[4, 2, 2]]$ inner code. (We did not compress the circuit as Ref. [11] did.) "7" indicates protocol using $[[7, 1, 3]]$ inner code. "15" is the Bravyi-Kitaev 15-to-1 protocol included for comparison purposes; "7" and "15" have almost identical performance. "16" is pipelined protocol using $[[16, 6, 4]]$ inner code. "20" is pipelined protocol using $[[16, 2, 4]]$ inner code. "17" is pipelined protocol using $[[17, 1, 5]]$ inner code. "22" is protocol using $[[21, 3, 5]]$ inner code with $n_{\text{outer}} = 4$. "21" is protocol using $[[21, 3, 5]]$ inner code pipelined with $[[7, 1, 3]]$ inner code. "23" is protocol using $[[23, 1, 7]]$ inner code pipelined with "21."

5.3.3 $[[23, 1, 7]]$ Inner Code

In the appendix, we give a $[[23, 1, 7]]$ inner code. Pipelining this code with a $[[7, 1, 3]]$ and a $[[17, 1, 5]]$ inner code gives us $n_{\text{outer}} = 1$ with error $O(\epsilon^7)$. This protocol consumes $1 + 7 \times 2 + 17 \times 2 + 23 \times 2 = 95$ T's, to produce 1 output magic state. We could also apply this code to each of the output bits of any of the other fifth order protocols of section 5.3.2 to obtain error $O(\epsilon^7)$; we do not show results for this here.

5.4 Results

The results of the simulations are shown in Fig. 5.1. Note that the plots are close to linear on a log-log plot, with only small deviations at high error rate. Each data point represents the average of at least 10^4 runs, with statistical flucutations negligible on the scale of the plot. The

asymptotic behavior is within statistical error of that given by an enumeration of minimum weight error patterns.

The protocol using $[[16, 2, 4]]$ inner code has a slightly reduced output error, compared to the protocol using the $[[16, 6, 4]]$ inner code.

We emphasize that ϵ_{out} indicates the probability that there is any error in the output state which is a multi-qubit state. Suppose that two protocols give the same value of ϵ_{out} for a given ϵ_{in}, but one protocol has a large n_{outer}. If the total number of magic states needed in our computation is large compared to n_{outer}, the number of times we need to call the protocol is inversely proportional to n_{outer}, and so the protocol with the larger n_{outer} for the given ϵ_{out} is less likely to produce an error.

The probability that no error is detected by the protocol is roughly $(1-\epsilon_{in})^{n_T}$. This result would be exact if any error in an input T gate led to the protocol detecting an error. Instead, some high weight error patterns do not lead to any error detected by the code, leading to slight corrections to this formula.

6 Discussion

We have given a general scheme to construct distillation protocols using inner and outer codes. If desired, our protocols can be concatenated with other protocols. However, on their own, they achieve asymptotic behavior which is conjectured to be optimal, as well as having small size examples which perform well. The concatenation can be useful in an architecture where Clifford gates are performed at a logical level so that one can tune the fidelity of Clifford gates to match the relatively low fidelity of output magic states from early stages of concatenation [41–43]. In contrast, our schemes without concatenation can be useful if Clifford gates are already of high fidelity.

One of the major advantages of our protocols is the small number of qubits that they use, as they maintain a constant ratio of physical to logical qubits in the asymptotic limit. It is interesting to consider the asymptotics of this overhead between physical and logical qubits. Note that given any distillation protocol, there is a trivial way to define a new protocol with a fixed ratio of physical to logical qubits. Suppose, for example, that some protocol uses n_{phys} qubits to produce 1 output magic state. Here n_{phys} includes all physical qubits used in the protocol, not only initial noisy T states. Call this protocol P. One can define a new protocol P' that works on $2n_{\text{phys}}$ qubits to produce n_{phys} output magic states: First, by working on n_{phys} qubits, leaving other n_{phys} inactive, we obtain 1 output, which is put away. Next, we use n_{phys} qubits out of $2n_{\text{phys}}-1$ qubits to obtain the second output, which is put away. Next, we use n_{phys} qubits out of $2n_{\text{phys}}-2$ qubits to obtain the third output, which is put away. Continuing, we apply P a total of n_{phys} times sequentially. However, the circuit depth of P' now is proportional to n_{phys} times the depth of P.

In contrast, in Theorem 4.1, we obtain d-th order error reduction at fixed ratio of physical to logical with a T-gate depth proportional to d. That is, the protocols we have constructed are space and time efficient in terms of T-gates. It should be noted that we have ignored all Clifford gates, and thus the T-efficiency claim is meaningful when T-gates are much more costly than Clifford gates.

Including Clifford gates, the entire circuit of Theorem 4.1 has depth $O(d^2 \log d)$ if geometric locality of the gate is ignored. Indeed, an n-qubit stabilizer code's encoding circuit can be

constructed in depth $O(n \log n)$. It is essentially Gauss elimination of a binary symplectic matrix. The Gauss elimination for a single column can be done by a circuit of depth $O(\log n)$, and hence the entire Gauss elimination can be done by a circuit of depth $O(n \log n)$. Since we are using a good family of codes, the circuit depth of an encoding circuit is $O(d \log d)$, and there are $O(d)$ encoding/decoding steps. The total gate count (spacetime) per output is $O(d^2 \log d)$.

Jones [13] constructed a family of protocols giving $\gamma \to 1$. (For the definition of γ, see the discussion below Theorem 4.1.) This protocol builds upon Knill's [8], and is in fact a subclass of ours. Implicitly in Ref. [13], the inner code is obtained by concatenating a $[[k+4, k, 2]]$ code ν times where k is even, yielding $[[k^\nu + 4(k^\nu - 1)/(k-1), k^\nu, 2^\nu]]$ normal weakly self-dual codes. (For the definition of normal and hyperbolic codes, see 3.) The outer code is a hypercubic grid with checks along coordinate axes. The dimension of the grid is proportional to $\nu \cdot 2^\nu$, since a check using an inner code of distance 2^ν has to be implemented $2^{\nu-1}$ times along $2^{\nu-1}$ independent axes, and the concatenation for the inner codes is also performed on a grid. For a given ν, in the large k limit, one obtains a protocol that consumes $2^\nu + 1 = d + 1$ noisy T gates per output at d-th order of reduction in error. The asymptotic performance of this family is similar to our Theorem 4.2. If we worked with normal codes of even distance in Theorem 4.2, then we would have concluded that the input T count per output is $d+1$. Note that the space requirement of Jones' scheme is much larger than that of Theorem 4.1, as the grid outer code used by Jones would require roughly $k^{\nu 2^\nu} = k^{O(d)}$ qubits. This exponential dependence on d also holds for the protocols in the proof of Theorem 4.2. In contrast, Theorem 4.1 requires only $O(d)$ qubits.

In comparison with Jones' scheme, our main technical contribution is to explicitly separate inner and outer codes with general criteria for them to be useful in distillation. The criteria are that inner codes have to be weakly self-dual, and outer codes have to be sensitive (i.e., the parity check matrix M should satisfy $2|Mv| + |v| \geq d$ for any nonzero error vector v). These requirements are rather simple, so we were able to consider random constructions for them. In fact, a large pool of existing codes can be incorporated. In particular, we note that there are quantum BCH codes [44], some of which have encoding rate greater than 1/2 with code distances 5 and 7 at modest code lengths.

For Theorem 4.2 we have resorted to a graph theoretic construction of outer codes from Ref. [40]. This is sufficient for the proof, but one may wish to have more concrete examples. In fact, a hypercubic grid of dimension $D \geq 3$ yields an outer code of desired sensitivity for $d = 2D + 1$, which will be analyzed in detail elsewhere [45].

References

[1] P. W. Shor, "Fault-tolerant quantum computation," in *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on* (1996) pp. 56–65, quant-ph/9605011 .

[2] D. Aharonov and M. Ben-Or, "Fault tolerant quantum computation with constant error," SIAM J. Comput. **38**, 1207–1282 (2008), quant-ph/9611025v2 .

[3] E. Knill, R. Laflamme, and W. Zurek, "Threshold accuracy for quantum computation," (1996), quant-ph/9610011 .

[4] D. Gottesman, "A class of quantum error-correcting codes saturating the quantum hamming bound," Phys. Rev. A **54**, 1862 (1996), quant-ph/9604038 .

[5] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane, "Quantum error correction and orthogonal geometry," Phys. Rev. Lett. **78**, 405–408 (1997), quant-ph/9605005 .
[6] N. C. Jones, R. V. Meter, A. G. Fowler, P. L. McMahon, J. Kim, T. D. Ladd, and Y. Yamamoto, "Layered architecture for quantum computing," Physical Review X **2**, 031007 (2012), 1010.5022 .
[7] J. O'Gorman and E. T. Campbell, "Quantum computation with realistic magic state factories," 1605.07197v2 .
[8] E. Knill, "Fault-tolerant postselected quantum computation: Schemes," (2004), quant-ph/0402171v1 .
[9] E. Knill, "Fault-tolerant postselected quantum computation: Threshold analysis," (2004), quant-ph/0404104v1 .
[10] S. Bravyi and A. Kitaev, "Universal quantum computation with ideal Clifford gates and noisy ancillas," Phys. Rev. A **71**, 022316 (2005), quant-ph/0403025 .
[11] A. M. Meier, B. Eastin, and E. Knill, "Magic-state distillation with the four-qubit code," Quant. Inf. Comp. **13**, 195 (2013), 1204.4221 .
[12] S. Bravyi and J. Haah, "Magic state distillation with low overhead," Phys. Rev. A **86**, 052329 (2012), 1209.2426 .
[13] C. Jones, "Multilevel distillation of magic states for quantum computing," Phys. Rev. A **87**, 042305 (2013), 1210.3388v2 .
[14] G. Duclos-Cianci and D. Poulin, "Reducing the quantum computing overhead with complex gate distillation," Phys. Rev. A **91**, 042315 (2015), 1403.5280v1 .
[15] I. L. Chuang and D. Gottesman, "Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations," Nature **402**, 390–393 (1999).
[16] A. J. Landahl and C. Cesare, "Complex instruction set computing architecture for performing accurate quantum Z rotations with less magic," (2013), 1302.3240 .
[17] A. Paetznick and B. W. Reichardt, "Universal fault-tolerant quantum computation with only transversal gates and error correction," Phys. Rev. Lett. **111**, 090505 (2013), 1304.3709 .
[18] T. Jochym-O'Connor and R. Laflamme, "Using concatenated quantum codes for universal fault-tolerant quantum gates," Phys. Rev. Lett. **112**, 010505 (2014), 1309.3310 .
[19] J. T. Anderson, G. Duclos-Cianci, and D. Poulin, "Fault-tolerant conversion between the steane and reed-muller quantum codes," Phys. Rev. Lett. **113**, 080501 (2014), 1403.2734 .
[20] H. Bombin, "Gauge color codes: Optimal transversal gates and gauge fixing in topological stabilizer codes," New J.Phys. **17**, 083002 (2015), 1311.0879v6 .
[21] S. Bravyi and A. Cross, "Doubled color codes," 1509.03239v1 .
[22] T. Jochym-O'Connor and S. D. Bartlett, "Stacked codes: Universal fault-tolerant quantum computation in a two-dimensional layout," Phys. Rev. A **93**, 022323 (2016), 1509.04255 .
[23] C. Jones, P. Brooks, and J. Harrington, "Gauge color codes in two dimensions," Phys. Rev. A **93**, 052332 (2016), 1512.04193 .
[24] H. Bombin, "Dimensional jump in quantum error correction," New Journal of Physics **18**, 043038 (2016), 1412.5079v3 .
[25] T. J. Yoder, R. Takagi, and I. L. Chuang, "Universal fault-tolerant gates on concatenated stabilizer codes," Physical Review X **6**, 031039 (2016), 1603.03948 .

[26] B. Eastin and E. Knill, "Restrictions on transversal encoded quantum gate sets," Phys. Rev. Lett. **102**, 110502 (2009), 0811.4262 .
[27] S. Bravyi and R. König, "Classification of topologically protected gates for local stabilizer codes," Phys. Rev. Lett **110**, 170503 (2013), 1206.1609 .
[28] E. T. Campbell and J. O'Gorman, "An efficient magic state approach to small angle rotations," Quantum Science and Technology **1**, 015007 (2016), 1603.04230v2 .
[29] B. W. Reichardt, "Improved magic states distillation for quantum universality," Quant. Inf. Proc. **4**, 251–264 (2005), quant-ph/0411036v1 .
[30] A. M. Steane, "Error correcting codes in quantum theory," Physical Review Letters **77**, 793–797 (1996).
[31] H. Bombin and M. A. Martin-Delgado, "Topological quantum distillation," Physical Review Letters **97**, 180501 (2006), quant-ph/0605138 .
[32] B. Eastin, "Distilling one-qubit magic states into toffoli states," Phys. Rev. A **87**, 032321 (2013), 1212.4872v2 .
[33] C. Jones, "Low-overhead constructions for the fault-tolerant toffoli gate," Physical Review A **87**, 022328 (2013), 1212.5069 .
[34] E. T. Campbell and M. Howard, "Unified framework for magic state distillation and multiqubit gate synthesis with reduced resource cost," Phys. Rev. A **95**, 022316 (2017), 1606.01904 .
[35] S. Bravyi, B. Leemhuis, and B. M. Terhal, "Majorana fermion codes," New J.Phys. **12**, 083039 (2010), 1004.3791 .
[36] A. R. Calderbank and P. W. Shor, "Good quantum error-correcting codes exist," Phys. Rev. A **54**, 1098–1105 (1996), quant-ph/9512032 .
[37] D. Gottesman, "Stabilizer codes and quantum error correction," (1997), quant-ph/9705052 .
[38] J. Preskill, "Lecture notes on quantum computation," Caltech Ph219 .
[39] M. Sipser and D. A. Spielman, "Expander codes," IEEE Transactions on Information Theory **42**, 1710–1722 (1996).
[40] Z. Furedi, F. Lazebnik, A. Seress, V. A. Ustimenko, and A. J. Woldar, "Graphs of prescribed girth and bi-degree," Journal of Combinatorial Theory, Series B **64**, 228–239 (1995).
[41] R. Raussendorf, J. Harrington, and K. Goyal, "Topological fault-tolerance in cluster state quantum computation," New Journal of Physics **9**, 199 (2007), quant-ph/0703143v1 .
[42] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, "Surface codes: Towards practical large-scale quantum computation," Phys. Rev. A **86**, 032324 (2012), 1208.0928 .
[43] J. O'Gorman and E. T. Campbell, "Quantum computation with realistic magic-state factories," Phys. Rev. A **95**, 032338 (2017), 1605.07197 .
[44] M. Grassl and T. Beth, "Quantum bch codes," in *Proceedings X. International Symposium on Theoretical Electrical Engineering, Magdeburg* (1999) pp. 207–212, quant-ph/9910060 .
[45] J. Haah, M. B. Hastings, D. Poulin, and D. Wecker, "Magic state distillation at intermediate size," 1709.02789 .
[46] M. B. Hastings, "Small majorana fermion codes," 1703.00612 .

[47] E. T. Campbell, H. Anwar, and D. E. Browne, "Magic state distillation in all prime dimensions using quantum reed-muller codes," Phys. Rev. X **2**, 041021 (2012), 1205.3104 .
[48] E. T. Campbell, "Enhanced fault-tolerant quantum computing in d-level systems," Phys. Rev. Lett. **113**, 230501 (2014), 1406.3055 .
[49] G. Nebe, E. M. Rains, and N. J. A. Sloane, "The invariants of the clifford groups," Designs, Codes and Cryptography **24**, 99–122 (2001), math/0001038 .
[50] G. Nebe, E. M. Rains, and N. J. A. Sloane, *Self-Dual Codes and Invariant Theory* (Springer-Verlag, 2006).
[51] E. F. Assmus and J. D. Key, *Designs and their Codes*, 103 (Cambridge University Press, 1992).
[52] S. Vijay and L. Fu, "Quantum error correction for complex and majorana fermion qubits," 1703.00459 .
[53] S. Lang, *Algebra*, revised 3rd ed. (Springer, 2002).

A Specific Small Inner and Outer Codes

In this appendix, we give some specific inner and outer codes, either giving the stabilizers or referring to the literature. Some of these codes are explained in the basic distillation section 1 or in numerical simulations 5 in the body of the paper. Other codes have other useful properties that we describe for specific codes.

When we give stabilizers for an inner code, each row gives one stabilizer generator. Each row consists of a binary string, of length equal to the number of qubits, with a 1 indicating that that stabilizer acts on that qubit, i.e., we give the parity check matrix.

A.1 Inner Codes

A.1.1 $[[4, 2, 2]]$ Inner Code

This is explained in Section 1.

A.1.2 $[[16, 6, 4]]$ Inner Code

This is from Ref. [46]. The stabilizer matrix is the classical Hadamard code $[16, 5, 8]$.

$$
\begin{array}{c}
1111111111111111 \\
1111111100000000 \\
1111000011110000 \\
1100110011001100 \\
1010101010101010
\end{array}
$$

A.1.3 $[[7, 1, 3]]$ Inner Code

This is explained in Section 1.

A.1.4 $[[17, 1, 5]]$ Inner Code

This is an instance of color code [21, 31]. It is the smallest normal code that we found with $k_{\text{inner}} = 1$ and distance 5. The stabilizers are:

$$
\begin{array}{c}
11011010101000010 \\
01100011001100110 \\
00110110010011001 \\
00010101000111110 \\
00001110010011101 \\
00000101000110000 \\
00000011111011010 \\
00000001010100001
\end{array}
$$

A.1.5 $[[21, 3, 5]]$ and $[[23, 1, 7]]$ Inner Codes

The (extended) Golay code is a classical self-dual code which has parameters $[24, 12, 8]$. Puncturing a bit by collecting all code words that has zero on that bit, we obtain a self-orthogonal $[23, 11, 7]$. From this, we obtain a weakly self-dual CSS code which is $[[23, 1, 7]]$. (Reichardt has used this code in a very different distillation protocol [29].) There are many positions to puncture, but due to high symmetry of the Golay code, the resulting codes have the same weight enumerators. One can pipeline the $[[23, 1, 7]]$ code after the protocol of section 1.1.2 to give a protocol with one output magic state and seventh order suppression in error.

By puncturing the $[[23, 1, 7]]$ code twice, we obtain a $[[21, 3, 5]]$ code. In a numerical search, we did not find any smaller normal code with $k_{\text{inner}} = 3$ and distance 5. The stabilizers of the $[[21, 3, 5]]$ code are:

$$
\begin{array}{c}
100000000011110110100 \\
010000000001111011010 \\
001000000110110011001 \\
000100000011011001101 \\
000010000001101100111 \\
000001000110111000110 \\
000000100101010010111 \\
000000010100100111110 \\
000000001100011101011
\end{array}
$$

A.1.6 Other Inner Codes

Some other examples of inner codes can be found in Ref. [46], from which we reproduce optimal k_{inner} found for given distance and n_{inner} in Table A.1. For stabilizers, see Ref. [46].

n_{inner}	16	20	24	28	30	20	28	30
k_{inner}	6	8	12	14	16	2	4	6
d	4	4	4	4	4	6	6	6

Table A.1: Parameters of small hyperbolic weakly self-dual CSS codes [46]. The code $[[20, 2, 6]]$ can be constructed from the five-qubit code $[[5, 1, 3]]$ by going through the Majorana operators [35], while the others cannot be constructed in this way.

A.2 Outer Codes

A.2.1 Petersen Graph Code

The outer code in section 5.3.2 has 4 qubits uses 4 checks of weight 3. However, from Lemma 4.15, we know that there is some n_{outer} such that there is a code which is $(4, 2)$ sensitive with weight-3 checks, which has only $(2/3)n_{\text{outer}}$ checks. We now explain this code. The proof of Lemma 4.15 reduces the problem of finding such a code to finding a bipartite graph G. Since the set B of that lemma has degree 2, we can equivalently define the code by a graph H such that the vertices of the graph H correspond to checks and the edges correspond to bits; i.e., in the case that B has degree 2, the possible bipartite graphs G are in one-to-one correspondence with degree-3 graphs H. Then, from the proof of Lemma 4.15 we know that if H has girth at least 5, then the corresponding code is $(4, 2)$ sensitive. The smallest such graph H is known to be the Petersen graph. This is a degree-3 graph with 15 vertices and 10 edges. This graph can be thought of as the dodecahedron with antipodes identified. Note that the girth being 5 is optimal in this case, because if H has girth 4, then there is a weight 4 error that violates no checks.

B Circuits

In this section we give circuits for some of the protocols above. Boxes labelled Enc or Enc' denote encoding and decoding circuits, which are Cliffords. The number in the box indicates what code is used. H denotes Haamard, M denotes measurement in Z basis, JM_x denotes measurement in X basis, Czs denotes control-Z operations.

C Coincidence among protocols

The Steane code has 7 Y-logical operators of weight 3. In the distillation protocol using the Steane code as the inner code, each logical error may appear in 4 different ways in the column that implements control-$H^{\otimes 7}$ [8]. The measurement error at the lowest order can happen in 7 ways. Overall, the cubic error can happen in $7 \cdot 4 + 7 = 35$ ways. This number matches the number of logical operators of weight 3 in the Bravyi-Kitaev 15-to-1 protocol [10]. Reichardt [29] has noted this equivalence.

When we pipelined $[[7, 1, 3]]$ to $[[17, 1, 5]]$, there are 48 T gates and 1 T states. The number of logical operators of weight 5 in $[[17, 1, 5]]$ is 51. Each logical operator can appear in 16 different configurations in the column that implements control-$H^{\otimes 17}$. The measurement error from the 17-qubit code routine occurs in 17 ways at the leading order. Thus, the output error

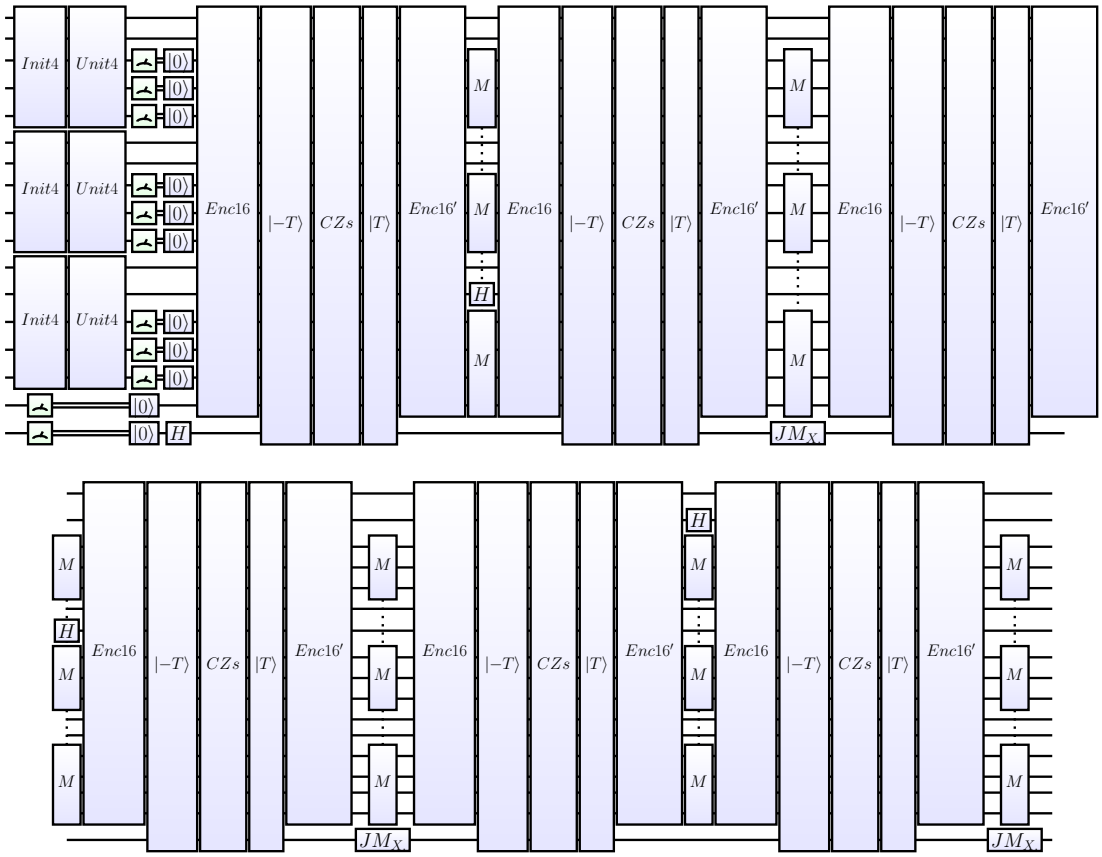


Figure B.1: Pipelined circuit using $[[16, 6, 4]]$ code described in section 1.2.2.

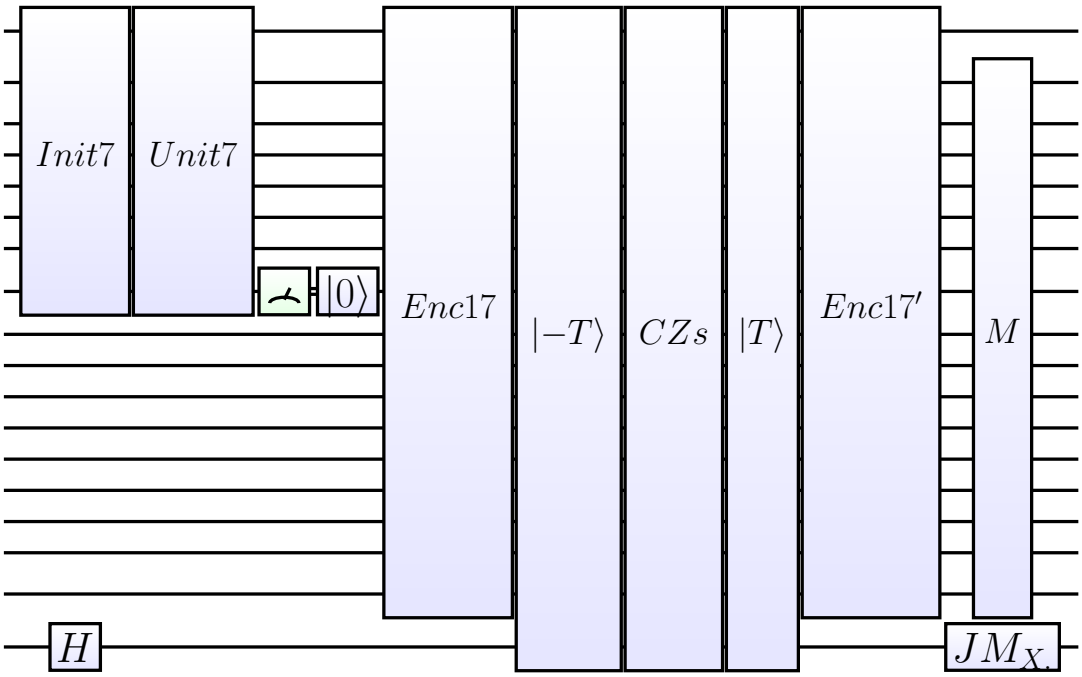


Figure B.2: Pipelined circuit using $[[17, 1, 5]]$ code described in section 1.1.2.

probability has leading term $(51 \cdot 16 + 35 \cdot 17)\epsilon^5 = 1411\epsilon^5$. The coefficient matches the number of Z-logical operators of weight 5 in the $[[49, 1, 5]]$ code, as reported in [12, App. B].

Bravyi and Cross [21] gave a recursive construction for triply even codes. They showed how to convert a pair of a (classical) triply even code of length n_{t-1} with dual distance $2t-1$ and some (classical) self-orthogonal code of length m_t with dual distance $2t+1$ into a triply even code of length $n_t = 2m_t + n_{t-1}$ with dual distance $2t+1$. The formula gives another coincidence with our pipeline. n_{t-1} is the number of T gates/states, sitting before the final H-measurement routine in the pipeline, and m_t is the code length of the final H-measurement routine. Thus, the recursive formula $n_t = 2m_t + n_{t-1}$ correctly counts the number of T gates/states used in the pipeline.

A similar coincidence was observed by Jones [13], where the leading error probabilities of the distillation protocols by a family of weakly self-dual $[[k+4, k, 2]]$ codes with $(k, 0)$-magic basis and those by a family of triorthogonal codes [12] are shown to be the same as $(3k+1)\epsilon^2$. The total number of T gates/states were also the same as $3k+8$.

D Qudits

In this section, we consider an extension to qudits with local Hilbert space dimension $p > 2$, with p a prime. Previously, Reed-Muller codes over prime fields were used [47, 48], but our approach is more efficient. In terms of the scaling exponent γ (see Sec. 4), previous schemes for a fixed p did not achieve $\gamma \to 1$, whereas our protocols below will.[9] Specifically, we consider

[9] In Ref. [48], it is shown that γ can be arbitrary close to 1 in the limit of large p.

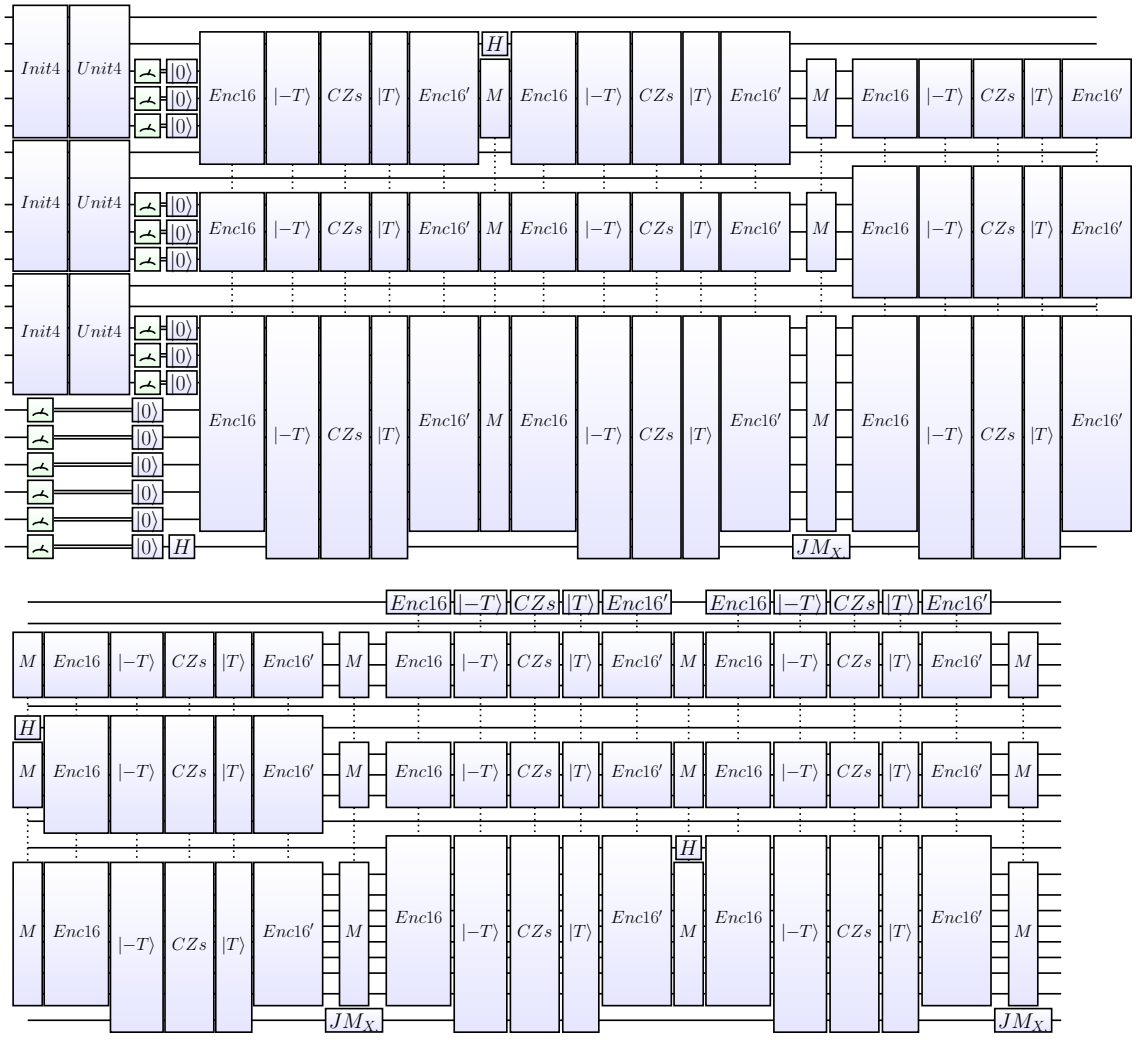


Figure B.3: Pipelined circuit using $[[16, 2, 4]]$ code described in section 5.3.1.

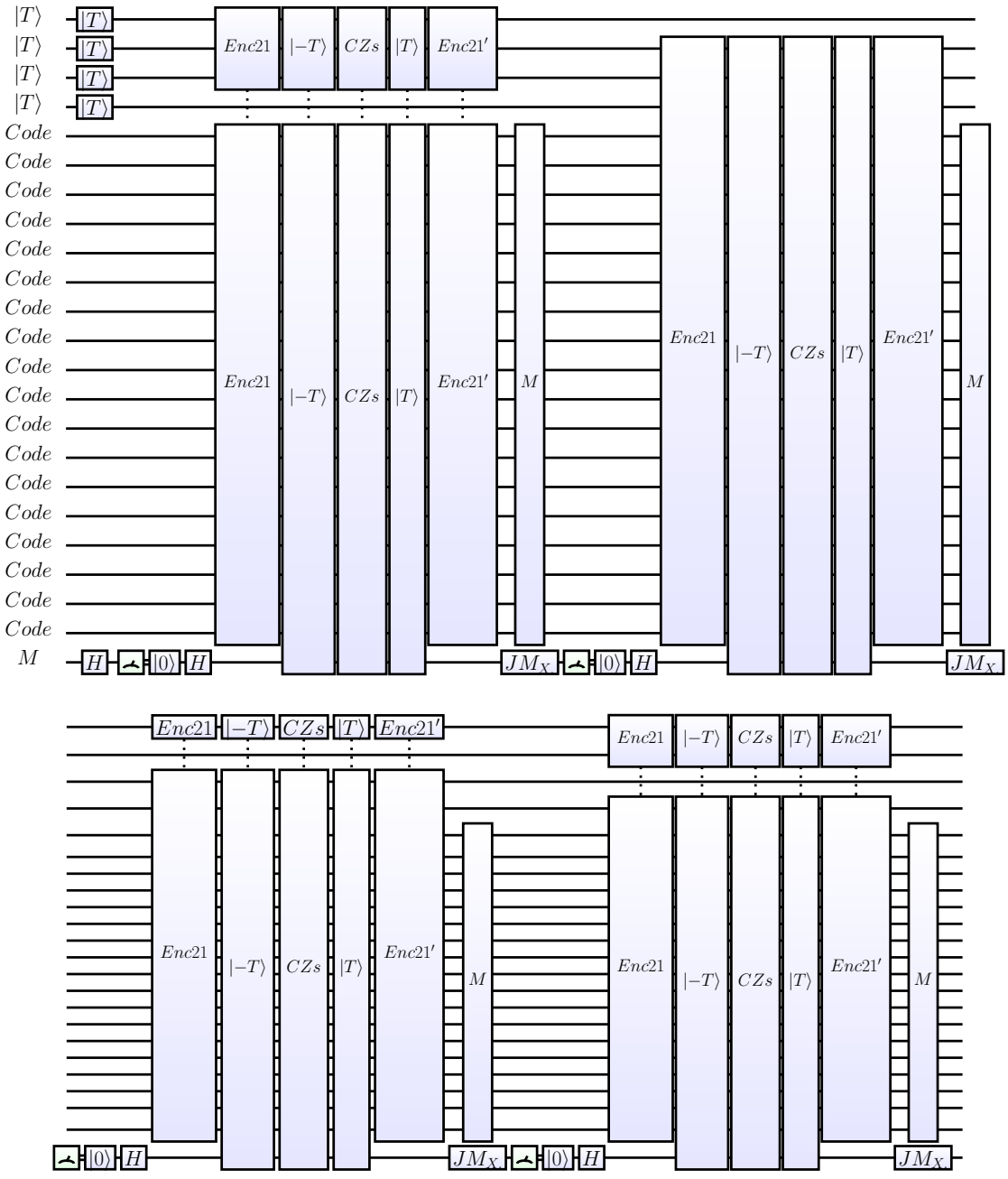


Figure B.4: Circuit using $[[21, 3, 5]]$ code described in section 5.3.2

outer codes defined by *binary* parity check matrices M that are sensitive enough. Those from biregular bipartite graph can be used. The parity check matrices being binary is for technical simplicity, and is not conceptually crucial ingredient, though relaxing this condition might involve complicated calculation. The main difference from the qubit protocols is in inner codes. We only consider analogs of normal codes, where transversal S gates become logical S gates. We show that measurement in the eigenbasis of SX can be implemented fault-tolerantly on an inner code of parameters $[[n_{\text{inner}}, k_{\text{inner}}, d]]_p$ using $4n_{\text{inner}}$ T gates if $p > 3$, or $2n_{\text{inner}}$ T gates if $p = 3$. Therefore, the Lemma 4.4 generalizes to odd prime dimensional qudits, where input T count is $n_{\text{outer}} + 4n_{\text{inner}}m$ if $p > 3$ where m is the number of checks in the outer code, or $n_{\text{outer}} + 2n_{\text{inner}}m$ if $p = 3$. We complement the construction in this section with a probabilistic existence proof for a good family of inner codes in the sense of Section 4; see Lemma E.4 and combine it with the proof of Lemma 4.10. We also point out that quantum Reed-Muller code gives a family whose encoding rate approaches 1 for a given distance. Hence, the statement of Theorem 4.1 remains unchanged in either case where $p = 3$ or $p > 3$, but that of Theorem 4.2 for $p > 3$ becomes that, for odd $d \geq 5$ the number of input magic states per output approaches $1 + 4(d-1)/2 = 2d - 1$ in the large code length limit. The input T count per output is still $d(1 + o(1))$ if $p = 3$. If one wishes to improve the asymptotic input count for $p > 3$, then one has to solve an equation analogous to (D.19).

Consider a basis of states $|j\rangle$, where $j = 0, 1, \ldots, p-1$ is periodic mod p. We use the following operators and phase factor

$$\omega = e^{2\pi i/p}, \qquad Z = \sum_j \omega^j |j\rangle\langle j|, \qquad X = \sum_j |j+1\rangle\langle j|,$$
$$H = \frac{1}{\sqrt{p}} \sum_{j,k} \omega^{jk} |j\rangle\langle k|, \qquad S = \sum_j \omega^{j(j-1)/2} |j\rangle\langle j|, \qquad {}^{C}X = \sum_j |j\rangle\langle j| \otimes X^j,$$
$$U(n) = \sum_j |nj\rangle\langle j| \quad (n \neq 0) \tag{D.1}$$

which generate the Clifford group. It holds that $ZX = \omega XZ$.

We will work with a generalization of *normal* codes throughout this section, ignoring hyperbolic codes. One reason is that we cannot achieve control-Swap in the same way as we could previously. The general method in the qubit case was to use some non-Clifford operation such as a T gate, conjugating controlled Pauli to obtain control-Swap on the code space of some code. However, Swap is of order 2 while control-Z is of order p. One might hope to obtain a control-permutation of order p, but we do not consider this possibility. For normal codes, we do not try to implement the control-Hadamard as was done before, because Hadamard is of order 4 for $p > 2$, and hence is not conjugate to control-Z.

D.1 Preliminary

Let us first define a T-gate [47]. The cases $p = 3$ and $p > 3$ are going to be different. Define

$$g(j) := \sum_{k=0}^{j} \frac{1}{2} k(k-1) = \frac{1}{6}(j^3 - j), \tag{D.2}$$
$$g(j+p) = g(j) \mod p \quad \text{if } p > 3 \tag{D.3}$$

where the second line is because 6 is invertible in $\mathbb{F}_{p>3}$, and ensures that g is a well defined function on $\mathbb{F}_{p>3}$. All arithmetic in the exponent of ω, Z, X, and S will be over $\mathbb{F}_p$ for both $p = 3$ and $p > 3$. Define the T-gate as

$$T = \sum_{j=0}^{p-1} \omega^{g(j)} |j\rangle \langle j| , \qquad TXT^{-1}X^{-1} = S \qquad \text{if } p > 3, \tag{D.4}$$

$$T = |0\rangle \langle 0| + e^{-2\pi i/9} |1\rangle \langle 1| + e^{2\pi i/9} |2\rangle \langle 2| , \qquad TXT^{-1}X^{-1} = e^{-2\pi i/9} S \qquad \text{if } p = 3. \tag{D.5}$$

These show that in both cases the T gate is at the third level of the generalized Clifford hierarchy. More generally, we find

$$T^m X T^{-m} = \begin{cases} S^m X & \text{for } p > 3, \\ e^{-2\pi i m/9} S^m X & \text{for } p = 3. \end{cases} \tag{D.6}$$

For both $p = 3$ and $p > 3$, define $|\psi_m\rangle$ for $m = 0, 1, \ldots, p-1$ be the $(+1)$-eigenstate of $T^m X T^{-m}$:

$$T^m X T^{-m} |\psi_m\rangle = |\psi_m\rangle . \tag{D.7}$$

Any state $|\psi_m\rangle$ for $m = 1, \ldots, p-1$ will be a "magic state."

How would one use these magic states? Suppose $p > 3$. Consider a pair of qudits in a state $\sum_j a_j |j\rangle \otimes |\psi_m\rangle$. Apply a control-$X$ operation with the first qudit as source and the second qudit as target. This maps the state to

$$\frac{1}{\sqrt{p}} \sum_{j,k} a_j \omega^{mg(k)} |j, k + j\rangle . \tag{D.8}$$

Now measure the second qudit in the computational basis, obtaining a result ℓ. This gives a state on the first qudit $\sum_j a_j \omega^{mg(\ell - j)} |j\rangle$. Thus, the transformation implemented on the first qudit is $\sum_j \omega^{mg(\ell-j)} |j\rangle \langle j|$. Expanding the exponent, we have

$$\begin{aligned} mg(\ell - j) &= mg(\ell) - mg(j) + \frac{m}{2}(\ell j^2 - \ell^2 j) \\ &= mg(\ell) - mg(j) + m\ell \frac{j(j-1)}{2} - m\frac{\ell(\ell-1)}{2} j. \end{aligned} \tag{D.9}$$

The first term on the right-hand side of Eq. (D.9) corresponds to an irrelevant global phase factor. The second term, $-mg(j)$, corresponding to implementing transformation T^{-m} on the first qudit. The third term gives a phase factor that can be corrected by applying a power of the S gate and the last term gives phase factors that can be corrected by a power of the Z gate. Thus, the state injection procedure works, in that we can use a magic state $|\psi_m\rangle$ to produce a transformation T^{-m} up to Clifford corrections.

When $p = 3$, we use the same state injection, with $m = 1$. One finds after some calculation that if the measurement outcome is $\ell = 0$, the implemented operations is T^{-1} to the source, if $\ell = 1$, it is $e^{-2\pi i/9} S T^{-1}$, and if $\ell = 2$, it is $e^{2\pi i/9} Z^{-1} S^{-1} T^{-1}$. Thus, in all cases, the implemented operation is T^{-1} up to a Clifford correction.

The injected T gates together with Cliffords form a universal gate set [47, App. D]. This is a corollary of [49, Thm. 7.3] that says the Clifford group is a maximal finite subgroup of

$U(p^n)$ up to global phase factors, and [50, Cor. 6.8.2] that says any infinite subgroup (even after quotienting out phase factors) containing the Clifford group is dense in $U(p^n)$.

Note that T^m and T^{-m} are interconvertible by Cliffords. More generally, it is possible to use Clifford operations to convert a gate T^m into another gate $T^{m'}$ with $m' = mn^3$ for $n \neq 0$, by $U = U(n) = \sum_j |nj\rangle\langle j|$ gate. For $p > 3$, we have $U^\dagger T^m U = \sum_j \omega^{mg(nj)} |j\rangle\langle j|$ where

$$g(nj) = \frac{1}{6}(n^3 j^3 - nj) = n^3 g(j) + \frac{n^3 - n}{6} j. \tag{D.10}$$

Thus $U^\dagger T^m U = T^{mn^3} Z^{m\frac{n^3-n}{6}}$, and so indeed $T^m = C_1 T^{m'} C_2$ for some Cliffords C_1, C_2. For $p = 3$, we see $T = U(-1)T^{-1}U(-1)$. Now, for which pairs m, m' can we find an n such that $m' = mn^3$? The multiplicative group $\mathbb{F}_p^\times$ is cyclic of order $p-1$. Therefore, when $p-1$ is not a multiple of 3, then $\mathbb{F}_p^\times \ni n \mapsto n^3 \in \mathbb{F}_p^\times$ is a bijection, and any T^m can be interconverted into any other $T^{m'}$. If $p-1$ is a multiple of 3, there are three distinct classes of T gates. Since $-1 = (-1)^3$, T^m and T^{-m} are always interconvertible.

D.2 Inner codes

For arbitrary vector $v \in \mathbb{F}_p^{n_{\text{inner}}}$ we write $X(v) = X^{v_1} \otimes \cdots \otimes X^{v_{n_{\text{inner}}}}$, and $Z(v) = Z^{v_1} \otimes \cdots \otimes Z^{v_{n_{\text{inner}}}}$. As in the weakly self-dual CSS code construction for qubits, it is straightforward to define a stabilizer code starting from a self-orthogonal subspace $\mathcal{S} \subset \mathcal{S}^\perp \subset \mathbb{F}_p^{n_{\text{inner}}}$: The stabilizer group is generated by $X(v)$ and $Z(v)$ where $v \in \mathcal{S}$. The quotient space $\mathcal{S}^\perp/\mathcal{S}$ is in one-to-one correspondence with the set of X-type (Z-type) logical operators, and the induced dot product on $\mathcal{S}^\perp/\mathcal{S}$ is non-degenerate. In Section E below, we show that there is a basis $\{v^{(1)}, \ldots, v^{(k_{\text{inner}})}\}$ of $\mathcal{S}^\perp/\mathcal{S}$ such that $v^{(i)} \cdot v^{(j)} = \alpha_j \delta_{ij}$ where the scalars α_j are all 1 possibly except the last one. For simplicity we restrict ourselves to cases where

$$(1, 1, \ldots, 1) \in \mathcal{S}, \tag{D.11}$$

$$v^{(i)} \cdot v^{(j)} = \delta_{ij}., \tag{D.12}$$

i.e., the second condition is that all scalars α_j are equal to 1. The first condition demands that n_{inner} to be a multiple of p. The second is a mild restriction, since $(\mathcal{S} \oplus \mathcal{S})^\perp/(\mathcal{S} \oplus \mathcal{S})$ always has a basis such that (D.12) holds. Given a basis $\{v^{(j)}\}$ satisfying (D.12), we define logical operators of the inner code as

$$\begin{aligned} \tilde{X}^{(j)} &= X(v^{(j)}), \\ \tilde{Z}^{(j)} &= Z(v^{(j)}), \end{aligned} \tag{D.13}$$

which indeed obey the commutation relation

$$\tilde{Z}^{(a)} \tilde{X}^{(b)} = \omega^{\delta_{ab}} \tilde{X}^{(b)} \tilde{Z}^{(a)} \tag{D.14}$$

of the generalized Pauli operators on k_{inner} qudits. Thus, this is a generalization of the normal codes in the qubit case. Due to (D.11), the transversal gate $\bar{S} = S^{\otimes n_{\text{inner}}}$ is a logical operator:

$$S^j X^k S^{-j} = \omega^{-jk(k+1)/2} Z^{jk} X^k \tag{D.15}$$

$$\bar{S} X(v) \bar{S}^{-1} = \omega^{-(v \cdot v + v \cdot \vec{1})/2} Z(v) X(v) \tag{D.16}$$

where in the second equation the phase factor vanishes when $v \in \mathcal{S}$.

We will implement the measurement of the stabilizer $T^m X T^{-m}$ of the magic state $|\psi_m\rangle$ using the inner codes. The measurement becomes feasible if ${}^C(T^m X T^{-m})$ can be implemented for *logical* qudits. We begin searching for its fault-tolerant implementation by observing an identity ${}^C(T^m X T^{-m}) = T^m({}^C X)T^{-m}$ that enables us to implement some controlled Clifford on logical qudits. The actual action on logical qubits depends on the inner code, but our conditions (D.11, D.12) will make it uniform across all logical qudits.

Recall $T^m X T^{-m} = \eta^{-1} S^m X$ where $\eta = 1$ if $p > 3$ and $\eta = e^{2\pi i/9}$ if $p = 3$. The action of the transversal gate $\bar{T}^m \bar{X} \bar{T}^{-m}$ can be deduced by looking at the logical operators and phase. The answer is

$$\bar{T}^m \bar{X} \bar{T}^{-m} = \eta^{-n_{\text{inner}}} \bar{S}^m \bar{X} \cong \eta^{k_{\text{inner}} - n_{\text{inner}}} \prod_{a=1}^{k_{\text{inner}}} \eta^{-1} (\tilde{S}^{(a)} (\tilde{Z}^{(a)})^{1/2})^m \tag{D.17}$$

because

$$\begin{cases} \bar{S}^m \tilde{X}^{(a)} \bar{S}^{-m} = \omega^{-m/2} (\tilde{Z}^{(a)})^m \tilde{X}^{(a)} \\ \bar{S}^m \tilde{Z}^{(a)} \bar{S}^{-m} = \tilde{Z}^{(a)} \\ \bar{S}^m \underbrace{\sum_{v \in \mathcal{S}} \frac{X(v)}{\sqrt{|\mathcal{S}|}} |0\rangle^{\otimes n_{\text{inner}}}}_{|\tilde{0}\rangle^{\otimes k_{\text{inner}}}} = |\tilde{0}\rangle^{\otimes k_{\text{inner}}} \end{cases} , \quad \begin{cases} (S^m Z^{m/2}) X (S^m Z^{m/2})^{-1} = \omega^{-m/2} Z^m X \\ (S^m Z^{m/2}) Z (S^m Z^{m/2})^{-1} = Z \\ (S^m Z^{m/2}) |0\rangle = |0\rangle \end{cases} . \tag{D.18}$$

Suppose $p > 3$. In order to implement ${}^C(\tilde{S}^m \tilde{X})$, we consider an equation and a solution

$${}^C(S^m X) = ({}^C X^{1-y})({}^C Z^u)({}^C S^x)({}^C X^y)({}^C S^z)({}^C Z^s)({}^C \omega^t) \tag{D.19}$$

$${}^C(S^m X) = ({}^C X^{1/3})[({}^C Z^{3m/8})({}^C S^{3m/4})]({}^C X^{2/3})[({}^C S^{m/4})({}^C Z^{m/8})]({}^C \omega^{-m/6}) \tag{D.20}$$

where the control is common for every gate, and u, x, y, z, s, t are variables. (Using ${}^C A = \sum_j |j\rangle\langle j| \otimes A^j$, one can evaluate matrix elements on both sides.) Note that the operators in the brackets are powers of ${}^C(S^m Z^{m/2})$. This implies that indeed simultaneous ${}^C(\tilde{S}^m \tilde{X})$ on all logical qudits can be implemented using $\bar{T}^m({}^C \bar{X}^{3m/4})\bar{T}^{-m}$, $\bar{T}^m({}^C \bar{X}^{m/4})\bar{T}^{-m}$, controlled Pauli logical operators, and a power of Z on the control.

When $p = 3$ it suffices to consider $m = 1$. To remove the phase factor $\eta^{k_{\text{inner}} - n_{\text{inner}}}$ we require that the k_{inner} is a multiple of 3. This can be achieved by considering three copies of a given code if necessary. n_{inner} is already a multiple of 3 due to (D.11). We can implement $\prod_{a=1}^{k_{\text{inner}}} {}^C(\eta^{-1} \tilde{S}^{(a)} \tilde{X}^{(a)})$ by an identity

$${}^C(\eta^{-1} SX) = ({}^C X^{-1})({}^C(\eta^{-1} SZ^{-1}))({}^C Z)({}^C X^{-1})({}^C Z^{-1})({}^C \omega^2). \tag{D.21}$$

We have shown that it is possible to build a fault-tolerant routine to measure $\tilde{T}^m \tilde{X} \tilde{T}^{-m}$.

We have not yet shown how to construct such inner codes. It is possible to generalize Lemma 4.9 to the case of matrices over a field $\mathbb{F}_p$ for $p > 2$; however, the generalization is more difficult since the self-orthogonality constraint implies a nonlinear constraint on the rows of the matrix so that each row is null; see Lemma E.4. Let us give an alternative construction which achieves the scaling similar to Lemma 4.11, namely that for any distance d, one can find

a family of normal weakly self-dual qudit CSS codes with $X(\vec{1})$ in the stabilizer group such that the ratio $k_{\text{inner}}/n_{\text{inner}} \to 1$ as $n_{\text{inner}} \to \infty$. This construction is derived from Reed-Muller codes. Let $C = RM_{\mathbb{F}_p}(r, m)$ be a classical Reed-Muller code over $\mathbb{F}_p$; the codewords have length p^m. The dual code is $C^\perp = RM_{\mathbb{F}_p}(m(p-1) - r - 1, m)$; see Theorem 5.4.2 of Ref. [51]. For any fixed r, for large enough m, $C \subset C^\perp$, so the codespace of C is self-orthogonal, and $\vec{1}$ is in the codespace of C. We use the codespace of C as the space $\mathcal{S}$, and use the CSS construction to define a weakly self-dual code. For fixed r, the rate of C tends to zero at large m, so the rate of the resulting weakly self-dual tends to 1. See Ref. [52] for weakly self-dual qubit codes derived from Reed-Muller codes. To make (D.12) hold, it may be necessary to use $\mathcal{S} \oplus \mathcal{S}$ instead of $\mathcal{S}$.

D.3 Outer codes

If the inner code has code distance d, then we should use an outer code with a parity check matrix that is $(d-1, \lceil \frac{d-1}{2} \rceil)$-sensitive. In full generality, one would want to use a parity check matrix with entries in $\mathbb{F}_p$, where an entry $\beta \neq 0$ would mean a stabilizer $(\eta^{-1} S^m X)^\beta$. This makes it necessary to have a different logical operator choice than we have used above.

However, a check matrix that is given by the adjacency matrix of a biregular graph with large girth is sufficient for us. Such a check matrix has only 0 and 1 entries, so no other choice of logical operator is necessary beyond what we have given above. Recall that a graph with large girth is locally a tree. Hence, a bad magic state will be caught by many checks because it flips a single stabilizer in these checks, and the required sensitivity is guaranteed.

E Symmetric forms over finite fields

We have classified nondegenerate symmetric forms over the binary field $\mathbb{F}_2$ in Section 3. Over a field of odd characteristic, the set of all finite dimensional vector spaces with nondegenerate symmetric forms (*quadratic spaces* for short) constitute an abelian group under the direct sum, after identifying hyperbolic planes as the identity. This group is known as the Witt group of the field, and the group structure is well known. Here we present a self-contained and elementary treatment of the Witt group of $\mathbb{F}_p$, and classify the quadratic spaces over fields of odd characteristic. A **square element**, or a **square** for short, is any member of the set $\{x^2 : x \in \mathbb{F}_p^2\}$.

It is natural to distinguish two cases depending on whether $-1 \in \mathbb{F}_p$ is a square, since a one-dimensional quadratic space is classified by $\mathbb{F}_p^\times / (\mathbb{F}_p^\times)^2$, where $\mathbb{F}_p^\times := \mathbb{F}_p \setminus \{0\}$ and $(\mathbb{F}_p^\times)^2 := \{x^2 \mid x \in \mathbb{F}^\times\}$. Since the multiplicative group $\mathbb{F}_p^\times$ is a cyclic group of order $p-1$, the element -1 being the unique element of $\mathbb{F}^\times$ with multiplicative order 2, is a square if and only if $p = 1 \mod 4$.

The part of the argument in Section 3 applies here without any change where we have inductively converted any non-degenerate symmetric matrix to a direct sum of a diagonal matrix and blocks of $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which represents a *hyperbolic plane*. Below, we assume that symmetric matrices are block diagonal in this form. It is then easy to explain why quadratic

spaces constitute a group:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^T \begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 2a \\ 2a & 0 \end{pmatrix} \simeq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{E.1}$$

This means that the one-dimensional quadratic space with form $(-a)$ is the inverse of the space with form (a). It is important here that 2 is an invertible element of the field.

We note that the determinant of the symmetric form up to squares is a nontrivial invariant valued in the multiplicative group $\mathbb{F}_p^\times/(\mathbb{F}_p^\times)^2$ which is isomorphic to the additive group $\mathbb{Z}/2\mathbb{Z}$. Let $\alpha \in \mathbb{F}_p$ be a non-square.

Case I: $p = 1 \mod 4$ so that $-1 \in (\mathbb{F}_p^\times)^2$. Consider a block $\mathrm{diag}(a, a)$ of the symmetric matrix. Since -1 is a square, we see $\mathrm{diag}(a, a) \simeq \mathrm{diag}(a, -a) \simeq \mathrm{diag}(1, -1) \simeq \mathrm{diag}(1, 1)$ under congruent transformations. Therefore, there are four classes of symmetric matrices up to hyperbolic planes: $\mathrm{diag}(1)$, $\mathrm{diag}(\alpha)$, $\mathrm{diag}(1, \alpha)$, and $\mathrm{diag}(1, 1)$. By looking at the determinant of the form and the parity of the dimension, we see that the four classes are distinct elements of the Witt group, which is hence isomorphic to $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. Given a dimension of quadratic spaces, we see there are only two exclusive possibilities:

$$\mathrm{diag}(1, 1, \ldots, 1, 1), \quad \text{and} \quad \mathrm{diag}(1, 1, \ldots, 1, \alpha). \tag{E.2}$$

Case II: $p = 3 \mod 4$ so that $-1 \notin (\mathbb{F}_p^\times)^2$. In this case, we can set $\alpha = -1$. We claim that $\mathrm{diag}(1, 1)$ is not hyperbolic. If $v = av_1 + bv_2$ is a vector in this two-dimensional space, where v_1, v_2 are basis vectors with $v_i^2 = 1$ and $a, b \in \mathbb{F}_p$, then $v \cdot v = a^2 + b^2$. Since -1 is not a square, the equation $a^2 + b^2 = 0$ does not have any nonzero solution, and this proves the claim. Next, we show that $\mathrm{diag}(1, 1) \simeq \mathrm{diag}(-1, -1)$. To this end, we will find a solution to $a^2 + b^2 + 1 = 0$ over $\mathbb{F}_p$. Once we have such a solution, then we see

$$\begin{pmatrix} a & b \\ b & -a \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ b & -a \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{E.3}$$

The existence of the solution follows from $(\mathbb{F}_p^\times)^2 + (\mathbb{F}_p^\times)^2 \not\subseteq (\mathbb{F}_p^\times)^2$, which implies that $(\mathbb{F}_p^\times)^2 + (\mathbb{F}_p^\times)^2 \ni -1$. If $(\mathbb{F}_p^\times)^2 + (\mathbb{F}_p^\times)^2 \subseteq (\mathbb{F}_p^\times)^2$, then $(\mathbb{F}_p^\times)^2$ would be a monoid under addition contained in a finite group, and hence would be a group itself, which must contain $0 \notin (\mathbb{F}_p^\times)^2$. Therefore, quadratic spaces given a dimension are classified by the determinant of the form up to squares.

$$\mathrm{diag}(1, 1, \ldots, 1, 1), \quad \text{and} \quad \mathrm{diag}(1, 1, \ldots, 1, -1) \tag{E.4}$$

The Witt group of $\mathbb{F}_p$ is isomorphic to $\mathbb{Z}/4\mathbb{Z}$ generated by $\mathrm{diag}(1)$.

Now we state and prove some facts about quadratic spaces.

Lemma E.1 (Chapter XV Theorem 10.2 of Ref. [53])**.** *Let Q be a nondegenerate quadratic space. If two subspaces V and U are isomorphic by an isometry $\sigma : V \to U$, then there exists an isometry $\bar{\sigma} : Q \to Q$ such that $\bar{\sigma}|_V = \sigma$.*

Lemma E.2. *Let N be a null subspace (on which the symmetric form vanishes) of a nondegenerate quadratic space Q over $\mathbb{F}_p$. Then, Q is isometric to the orthogonal sum of $N^\perp/N$ and a minimal hyperbolic subspace that contains N.*

Proof. Applying Lemma E.1 to the identity map σ, we conclude that any orthogonal set of vectors extends to an orthogonal basis. Since the form is nondegenerate, there exists a minimal hyperbolic subspace that includes N (*hyperbolic extension*), and the symmetric form can be written as $\Lambda' \oplus \lambda$. where λ is hyperbolic, and Λ' is nondegenerate. It is then clear that $N^\perp/N$ has the symmetric form Λ'. □

Lemma E.3. *Let Q be a nondegenerate quadratic space of dimension n over $\mathbb{F}_p$. Every maximal null subspace of Q has the same dimension m. Given any null subspace N of dimension $k \le m$, the number of null vectors of Q that are orthogonal to N is*

$$\#\mathcal{Z}(Q,N) = p^{n-k-1} + p^m - p^{n-m-1} =: \zeta(n,m,k). \tag{E.5}$$

Proof. To prove the first claim, suppose M, M' are maximal null subspaces. If $\dim M \le \dim M'$, then any injection from M to M' is an isometry, which can be extended to Q as $\bar{\sigma}$. Then, $\bar{\sigma}^{-1}(M')$ is a null superset of M, and hence is M itself since M is maximal. Thus, $\dim M = \dim M'$.

Let $\mathcal{Z}(Q,N)$ be the set of all null vectors of Q that are orthogonal to N. ($\mathcal{Z}$ is not a subspace in general.) Consider $\phi : \mathcal{Z}(Q,N) \to \mathcal{Z}(N^\perp/N, 0)$, a restriction of the canonical projection map $Q \to Q/N$. The map ϕ is surjective by definition of ζ. If $x, y \in \mathcal{Z}(Q,N)$ are mapped to the same element, then $x - y \in N$. This implies that ϕ maps exactly $\#N$ elements to one. (Here, # denotes the number of elements of the finite set.) Therefore,

$$\#\mathcal{Z}(Q,N) = (\#N)(\#\mathcal{Z}(N^\perp/N, 0)). \tag{E.6}$$

Due to the preceding lemma, the dimension of a maximal null subspace of $N^\perp/N$ is $m-k$. Thus, it remains only to prove the lemma when $k = 0$ since

$$\#\mathcal{Z}(Q,k) = p^k(p^{n-2k-1} - p^{n-m-k-1} + p^{m-k}) = p^{n-k-1} - p^{n-m-1} + p^m. \tag{E.7}$$

A *definite* quadratic space is one in which $w \cdot w = 0$ implies $w = 0$.[10] To count all null vectors, we work in a basis such that the n-by-n symmetric matrix is

$$\Lambda = \Lambda' \oplus \Lambda_{2m} \tag{E.8}$$

where Λ' definite, and $\Lambda_{2m} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes I_m$ is an orthogonal sum of m hyperbolic planes. In this basis, let us write any vector x as $x' \oplus (u, u')$. The nullity is then expressed by a quadratic equation of coordinates

$$x' \cdot x' + u \cdot u' = 0. \tag{E.9}$$

The solutions of this equation are divided into two classes: $x' \cdot x' = 0$ or $x' \cdot x' \neq 0$. In the former case, $x' = 0$ and $u \cdot u' = 0$. Given arbitrary u there is u' such that this equation holds. The number of solutions is $p^m + (p^m - 1)p^{m-1}$. In the latter case, we must have $u \neq 0$, and $u \cdot u' = c = -x' \cdot x' \neq 0$ is a inhomogeneous equation in u', whose solution always exists. For any given nonzero c, there are thus $(p^m - 1)p^{m-1}$ choices of (u, u'). x' can be any nonzero

[10] A definite space corresponds to a nontrivial element of the Witt group. For example, a one-dimensional nondegenerate space is definite. If α is not a square, then the symmetric form $\mathrm{diag}(1, -\alpha)$ is definite.

vector, so there are $p^{n-2m}-1$ choices. In sum, the number of null vectors in an n-dimensional quadratic space Q over $\mathbb{F}_p$ is

$$\#\mathcal{Z}(Q,0) = p^m + (p^m-1)p^{m-1} + (p^{n-2m}-1)(p^m-1)p^{m-1} = p^{n-1} - p^{n-m-1} + p^m. \tag{E.10}$$

□

Lemma E.4. *Let $w_1 = \vec{1} \in \mathbb{F}_p^n$ be the all-1 vector where n is a multiple of $p \geq 3$. Assume $c < (n-2)/2$, and let $w_2, \ldots, w_c$ be null vectors of $\mathbb{F}_p^n$ chosen inductively such that w_j is chosen uniformly at random from $\mathcal{Z}(\mathbb{F}_p^n, V_{j-1})$ where $V_{j-1} = \operatorname{span}(w_1, \ldots, w_{j-1})$. Let M be a c-by-n matrix with rows w_j.*

Consider a fixed n-component vector v, with $v \neq 0$ and $v \neq \vec{1}$. The probability that $Mv = 0$ is bounded from above by

$$20\left(\frac{3}{5}\right)^{n-c} + \left(\frac{11}{15}\right)^{c-1}. \tag{E.11}$$

Proof. We will estimate the desired probability by a union bound, considering separately the event that $v \in V_c^\perp$ and $v \notin V_c$, and the event that $v \in V_c^\perp$ and $v \in V_c$. The second event is possible only if $v \cdot v = 0$. By the classification of symmetric forms, a maximal null space of $\mathbb{F}_p^n$ has dimension m such that $n-2 \leq 2m \leq n$. The assumption that $c < (n-2)/2$ implies that

$$k_j := \dim V_j \leq j \leq c \leq m-1. \tag{E.12}$$

Consider the first event, assuming $v \cdot v = 0$. Let $j > 1$. Then

$$\Pr[v \in V_j^\perp \text{ and } v \notin V_j | v \in V_{j-1}^\perp \setminus V_{j-1}] \leq \frac{\zeta(n,m,k_{j-1}+1)}{\zeta(n,m,k_{j-1})} \leq \frac{p}{2p-1} \leq \frac{3}{5} \tag{E.13}$$

because w_j has to be orthogonal to $\operatorname{span}(v) + V_{j-1}$, which is null and is a proper superset of V_{j-1}. Thus, for any t,

$$\begin{aligned}\Pr[v \in V_t^\perp \text{ and } v \notin V_t] &= \Pr[v \in V_1^\perp \setminus V_1] \prod_{j=2}^{t} \Pr[v \in V_j^\perp \setminus V_j \mid v \in V_{j-1}^\perp \setminus V_{j-1}] \\ &\leq \prod_{j=2}^{t} \frac{\zeta(n,m,k_{j-1}+1)}{\zeta(n,m,k_{j-1})} \\ &\leq \left(\frac{3}{5}\right)^{t-1}. \end{aligned} \tag{E.14}$$

For $t = c$, we find in particular that

$$\Pr[v \in V_c^\perp \text{ and } v \notin V_c] \leq \left(\frac{3}{5}\right)^{c-1}. \tag{E.15}$$

Now assume $v \cdot v \neq 0$. The event that $v \in V_j^\perp$ happens only if w_j is chosen from $v^\perp$. We bound the decomposition $\Pr[v \in V_c^\perp] = \Pr[v \in V_1^\perp] \prod_{j=2}^{c} \Pr[v \in V_j^\perp | v \in V_{j-1}^\perp]$. The first term is bounded by 1 trivially. For other factors, we observe that the dimension of $v^\perp$ is $n-1$, and

a maximal null subspace in $v^\perp$ has dimension $m' \le m$. Under the conditioning $v \in V_{j-1}^\perp$, the null space V_{j-1} is a subspace of $v^\perp$, and $\#\mathcal{Z}(v^\perp, V_{j-1}) = \zeta(n-1, m', k_{j-1}) \le \zeta(n-1, m, k_{j-1})$. Hence,

$$\begin{aligned}\Pr[v \in V_c^\perp] &= \Pr[v \in V_1^\perp] \prod_{j=2}^{c} \Pr[v \in V_j^\perp | v \in V_{j-1}^\perp] \\ &\le \prod_{j=2}^{c} \frac{\zeta(n-1, m, k_{j-1})}{\zeta(n, m, k_{j-1})} \\ &\le \left(\frac{p^2+p-1}{2p^2-p}\right)^{c-1} \le \left(\frac{11}{15}\right)^{c-1} \end{aligned} \tag{E.16}$$

where in the second inequality, we used the assumption that $k_{j-1} \le c < (n-2)/2 \le m$.

Let us turn to the second event, assuming $v \cdot v = 0$. Note that if $v \in V_c$, there is a least j such that $v \in V_j$. So,

$$\Pr[v \in V_c^\perp \text{ and } v \in V_c] \le \sum_{j=2}^{c} \Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}]. \tag{E.17}$$

We have

$$\begin{aligned}&\Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \\ &= \Pr[v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \cdot \Pr[v \in V_j | v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \\ &\le \left(\frac{3}{5}\right)^{j-2} \Pr[v \in V_j | v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}], \end{aligned} \tag{E.18}$$

where we used Eq. (E.14). The second factor is bounded as

$$\Pr[v \in V_j | v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \le \frac{p^{1+k_{j-1}}}{\zeta(n, m, k_{j-1})} < \frac{3}{2p^{n-2j}} \tag{E.19}$$

because w_j belongs to $\mathrm{span}(v) + V_{j-1}$. Hence, $\Pr[v \in V_j \text{ and } v \in V_{j-1}^\perp \text{ and } v \notin V_{j-1}] \le 5(\frac{3}{5})^{n-j}$. So by Eq. (E.17),

$$\Pr[v \in V_c^\perp \text{ and } v \in V_c] \le \sum_{j=2}^{c} 5\left(\frac{3}{5}\right)^{n-j} < 20\left(\frac{3}{5}\right)^{n-c}. \tag{E.20}$$

Summing the probabilities of (E.16) and (E.20), we conclude the proof. □