Between Shor and Steane: A Unifying Construction for Measuring Error Syndromes

Shilin Huang[1,*] and Kenneth R. Brown[1,2,3,†]
[1]Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA
[2]Department of Physics, Duke University, Durham, North Carolina 27708, USA
[3]Department of Chemistry, Duke University, Durham, North Carolina 27708, USA

(Received 15 February 2021; accepted 22 June 2021; published 25 August 2021)

Fault-tolerant quantum error correction requires the measurement of error syndromes in a way that minimizes correlated errors on the quantum data. Steane and Shor ancilla are two well-known methods for fault-tolerant syndrome extraction. In this Letter, we find a unifying construction that generates a family of ancilla blocks that interpolate between Shor and Steane. This family increases the complexity of ancilla construction in exchange for reducing the rounds of measurement required to fault tolerantly measure the error. We then apply this construction to the toric code of size $L \times L$ and find that blocks of size $m \times m$ can be used to decode errors in $O(L/m)$ rounds of measurements. Our method can be applied to any Calderbank-Shor-Steane code and presents a new direction for optimizing fault-tolerant quantum computation.

DOI: 10.1103/PhysRevLett.127.090505

Quantum error-correction codes [1–9] are a path toward to the low errors required for large-scale quantum computation. Error correction is performed conditional on the measurement outcomes of a set of code stabilizers, also known as the error syndrome. The syndrome extraction circuits need to be *fault tolerant* [10–17] as the act of measuring syndromes also introduces extra errors on the quantum system. The first fault-tolerant syndrome extraction scheme was proposed by Shor [10,18]. In Shor's scheme, each syndrome bit is extracted from the data qubits to a verified ancilla cat state by transversal 2-qubit gates. Transversal operations limit the error propagation and no high-weight correlated errors can occur on the data qubits if the cat states are verified by postselection. For low-weight stabilizers, ancilla postselection can be avoided by decoding the ancilla cat states to look for potential correlated errors [15,19,20]. The value of the syndrome bit is the parity of the transversal measurement outcome of the cat state. As any measurement error will flip the syndrome bit, for a stabilizer code of distance d, one needs to repeat the syndrome measurements for $O(d^2)$ rounds to guarantee fault tolerance [10]. Optimizing the space and time overhead of Shor's scheme on particular codes is an active area of research with substantial progress since its invention [15,16,20–33].

The fault-tolerant extraction gadget for Shor's scheme is arguably the simplest. As a trade-off, a large number of 2-qubit gates are applied between data and ancilla qubits. For Calderbank-Shor-Steane (CSS) codes [5,6], Steane suggested to transfer the complexity of data-ancilla interaction to preparation of the ancilla state [12]. In Steane's protocol, a logical $|+\rangle$ ($|0\rangle$) ancilla state is prepared for simultaneously extracting all the Z stabilizers (X stabilizers). The errors on the data are propagated to the ancillas through transversal controlled-NOT (CNOT) gates. Steane's scheme requires no repetition of syndrome extractions, so that each data qubit is touched by 2 CNOT gates, one for X errors and another one for Z errors. Knill proposed a similar scheme [14] based on quantum teleportation that works for arbitrary stabilizer codes and requires only one round of transversal CNOT gates to extract all the stabilizers, at the cost of entangled logical ancilla. Using a constant number of Steane or Knill syndrome extractions, an arbitrary logical Clifford circuit can be implemented fault tolerantly in $O(1)$ steps [34].

For quantum devices with low idling error rates, the data qubits can wait for a reasonable amount of time until a good Steane-style (Knill-style) ancilla block is postselected, which yields higher fault-tolerance thresholds [14,35]. For large code blocks, however, postselection often yields an impractically low success rate. In this Letter, we develop a framework that generates a family of extraction schemes for CSS codes, including Shor's and Steane's construction as its two extremes. This family gradually increases the complexity of ancilla construction in exchange for reducing the number of 2-qubit gates between data and ancilla qubits required to fault tolerantly measure the error. As an example, we are able to use a single ancilla block to measure the plaquette operators (Z-stabilizer elements) of the toric code inside any connected sublattice. In particular, one can partition the $L \times L$ toric lattice into patches, each of which contains $m \times m$ plaquettes. Choosing m to be a constant independent of L, the success rate of ancilla postselection will be finite. Moreover, by offsetting the partition periodically, one can achieve fault tolerance within $O(L/m)$ measurement rounds. As a remark, our result is

compatible with the fact that Shor's and Steane's schemes require $O(L)$ and $O(1)$ measurement rounds on the toric code, respectively.

For CSS codes, the bit-flip errors (X errors) and phase-flip errors (Z errors) can be handled independently. Without loss of generality, we focus on bit-flip errors, which are detected by measuring Z checks. Suppose we are measuring r Z checks of an n-qubit CSS code. These checks can be represented by an $r \times n$ parity-check matrix H, where $H_{ij} = 1$ if and only if the ith Z check is supported on the jth qubit. The bit-flip errors on the physical qubits, represented by an n-bit binary string $\psi \in \mathbb{F}_2^n$, should have a syndrome $H\psi \in \mathbb{F}_2^r$.

In the intermediate steps of fault-tolerant computations, we are not allowed to measure the n data qubits directly. A general idea shared by both Shor's and Steane's syndrome extraction protocols is to transfer the X errors to a set of m ancilla qubits by CNOT gates: one can perform CNOT gates with data qubits as controls and ancilla qubits as targets, then apply Z measurements on all ancilla qubits to obtain the syndrome. The CNOT gates can be encoded by an $m \times n$ binary matrix Γ, where $\Gamma_{ij} = 1$ if and only if there is a CNOT gate between the jth data qubit and the ith ancilla qubit. For an error configuration $\psi \in \mathbb{F}_2^n$, the error transferred to the ancilla block is $\Gamma\psi \in \mathbb{F}_2^m$. The matrix Γ is referred to as the "gate matrix." To extract the syndrome, the ancilla block must be stabilized by r Z checks represented by an $r \times m$ matrix $\tilde{H}$ such that $H = \tilde{H}\Gamma$. This guarantees that one can obtain the syndrome $H\psi = \tilde{H}\Gamma\psi$ by measuring all the ancilla qubits in the Z basis. As an example, Steane's method corresponds to a decomposition $H = \tilde{H}\Gamma$ such that $\tilde{H} = H$ and $\Gamma = \mathbb{I}$ is the identity matrix. The case that $\Gamma = H$ and $\tilde{H} = \mathbb{I}$ corresponds to the non-fault-tolerant version of Shor's scheme, in which each stabilizer is measured by a single ancilla qubit.

A natural question is that, given an arbitrary matrix decomposition $H = \tilde{H}\Gamma$, can we find an extraction circuit with a gate matrix Γ? To obtain a deterministic syndrome $H\psi = \tilde{H}\Gamma\psi$, the ancilla state should at least be stabilized by a Z-stabilizer group isomorphic to $\operatorname{im}\tilde{H}^{\mathsf{T}}$, the image of $\tilde{H}^{\mathsf{T}}$ (the row space of $\tilde{H}$). On the other hand, we want to have as many X-stabilizer elements as possible to guarantee that Γ encodes a trivial logical operation and does not entangle the data and ancilla. The maximal X-stabilizer group we can have is isomorphic to $\ker\tilde{H}$, the kernel of $\tilde{H}$. In fact, our question has a positive answer:

Theorem 1.—A decomposition $H = \tilde{H}\Gamma$ uniquely corresponds to a syndrome extraction circuit with gate matrix Γ (up to gate ordering). The ancilla block is a CSS state whose Z-stabilizer and X-stabilizer groups are isomorphic to the row space of $\tilde{H}$, $\operatorname{im}\tilde{H}^{\mathsf{T}}$ and the kernel of $\tilde{H}$, $\ker\tilde{H}$, respectively.

The essential step to prove Theorem 1 is to verify that $\Gamma^{\mathsf{T}}(\operatorname{im}\tilde{H}^{\mathsf{T}}) \subseteq \operatorname{im}H^{\mathsf{T}}$ and $\Gamma(\ker H) \subseteq \ker\tilde{H}$ so that the stabilizer group and logical operators are invariant under Γ. A more comprehensive argument is presented in the companion article [36]. We present five methods of syndrome measurement for Steane's $[\![7,1,3]\!]$ code in Table I to show how our notation describes both standard methods for measuring syndromes and enables the discovery of new schemes. Circuits for these methods are presented in Ref. [36]. We note that nontransversal syndrome extraction circuits, such as the bare ancilla and scheme B circuits in Table I, require extra machinery to

TABLE I. For the Steane $[\![7,1,3]\!]$ code, we illustrate by matrices how our division of H into $\tilde{H}$ and Γ enables us to describe three common ancilla blocks: bare qubits, cat states, and Steane ancilla. It also enables protocols that extract two Z-stabilizer elements in parallel, as shown in schemes A and B. The Z- and X-stabilizer generators are generated from the image of $\tilde{H}^{\mathsf{T}}$, $\operatorname{im}\tilde{H}^{\mathsf{T}}$, and the kernel of $\tilde{H}$, $\ker\tilde{H}$, respectively. The circuits labeled bare ancilla and scheme B are not transversal so that an ancilla error could lead to a weight-2 correlated error on the data block. We will require flag qubits [15] or DiVincenzo-Aliferis decoding circuits [19] to detect these errors so that all single errors are distinguishable. For methods other than the Steane scheme, we show circuits that measure only a subset of the Z checks. To completely measure the syndrome, we can apply the same gadgets to other subsets or even combine schemes.

	Bare ancilla	Cat state	Steane	Scheme A	Scheme B
$\tilde{H}$	$\begin{bmatrix}1\end{bmatrix}$	$\begin{bmatrix}1&1&1&1\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0&0\\0&1&1&0&1&1&0\\0&0&1&1&0&1&1\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0\\0&1&1&0&1&1\end{bmatrix}$	$\begin{bmatrix}1&1&0\\0&1&1\end{bmatrix}$
Γ	$\begin{bmatrix}1&1&1&1&0&0&0\end{bmatrix}$	$\begin{bmatrix}\mathbb{I}_4 \mid 0_{4\times3}\end{bmatrix}$	$\mathbb{I}_7$	$\begin{bmatrix}\mathbb{I}_6 \mid 0_{6\times1}\end{bmatrix}$	$\begin{bmatrix}1&0&0&1&0&0&0\\0&1&1&0&0&0&0\\0&0&0&0&1&1&0\end{bmatrix}$
H	$\begin{bmatrix}1&1&1&1&0&0&0\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0&0\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0&0\\0&1&1&0&1&1&0\\0&0&1&1&0&1&1\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0&0\\0&1&1&0&1&1&0\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&0&0&0\\0&1&1&0&1&1&0\end{bmatrix}$
Z-stabilizer generators	$Z_{1'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}$, $Z_{2'}Z_{3'}Z_{5'}Z_{6'}$, $Z_{3'}Z_{4'}Z_{6'}Z_{7'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}$, $Z_{2'}Z_{3'}Z_{5'}Z_{6'}$	$Z_{1'}Z_{2'}$, $Z_{2'}Z_{3'}$
X-stabilizer generators	None	$X_{1'}X_{2'}$, $X_{2'}X_{3'}$, $X_{3'}X_{4'}$	$X_{1'}X_{2'}X_{3'}X_{4'}$, $X_{2'}X_{3'}X_{5'}X_{6'}$ $X_{3'}X_{4'}X_{6'}X_{7'}$, $X_{1'}X_{2'}X_{5'}$	$X_{1'}X_{2'}X_{3'}X_{4'}$, $X_{2'}X_{3'}X_{5'}X_{6'}$ $X_{3'}X_{4'}X_{6'}$, $X_{1'}X_{2'}X_{5'}$	$X_{1'}X_{2'}X_{3'}$

prevent correlated data errors such as flag qubits [15] or DiVincenzo-Aliferis decoding [19].

We now consider block extractions of the toric code. A $[\![2L^2, 2, L]\!]$ toric code defined on an $L \times L$ periodic lattice on the torus has a set V of L^2 vertices, a set E of $2L^2$ edges, and a set F of L^2 faces. The "boundary map" $\partial\colon \mathbb{F}_2[F] \to \mathbb{F}_2[E]$ is a $\mathbb{F}_2$-linear map that maps each face $f \in F$ to the sum of four edges bordering f. The "coboundary map" $\delta\colon \mathbb{F}_2[V] \to \mathbb{F}_2[E]$ maps each vertex $v \in V$ to the sum of four edges incident to v. The maps ∂^T and δ^T are the Z- and X-check matrices of the toric code, respectively. From Theorem 1, a circuit that extracts the Z stabilizers corresponds to a decomposition $\partial^\mathsf{T} = \tilde{\partial}^\mathsf{T}\gamma^\mathsf{T}$, or $\partial = \gamma\tilde{\partial}$ for some γ and $\tilde{\partial}$. We note that such decomposition can be constructed by cutting the torus along some arbitrarily chosen edges. The obtained topological space will have an edge set $\tilde{E}$ and a boundary map $\tilde{\partial}\colon \mathbb{F}_2[F] \to \mathbb{F}_2[\tilde{E}]$, and $\gamma\colon \mathbb{F}_2[\tilde{E}] \to \mathbb{F}_2[E]$ maps the edge $\tilde{e} \in \tilde{E}$ to the corresponding edge $e \in E$ on the torus. Note that, if we split an edge $e \in E$ into two edges $\tilde{e}_1, \tilde{e}_2 \in \tilde{E}$, $\tilde{e}_1, \tilde{e}_2$ will be on the boundary of the obtained topological space. As a remark, if the torus is not being cut at all, we obtain a Steane-style gadget; if the torus is cut into L^2 disjoint faces so that all the edges are split, we obtain the Shor-style gadget. Note that, for any extraction gadget obtained in this way, the 2-qubit gates are transversal on the ancilla block. Indeed, fault tolerance can be achieved when the ancilla block is prepared fault tolerantly. Such a family of "transversal gadgets" can be constructed algebraically on any CSS codes, which is shown in Ref. [36].

We now describe our model of fault-tolerant error correction. As X and Z errors can be corrected separately, we can ignore the Z errors on the data qubits and the X-stabilizer measurements. Suppose our computation starts from time 0. We extract the Z-check matrix ∂^T at every positive integer time. For every $t \in \mathbb{Z}_+$, each data qubit could have an X error at time $t - 1/2$, and the measurement outcome of each ancilla qubit at time t could have a classical bit-flip error. For now, we ignore the data errors between two CNOTs in the same extraction round. Suppose at time $t \in \mathbb{N}$, we apply a transversal gadget $\partial = \gamma_t\tilde{\partial}_t$ with an edge set $\tilde{E}_t$. If the syndrome bit $f \in F$ at time t is different from its previous value, we create a defect at the lattice site (f, t). A bit-flip error on the data qubit $e \in E$ at time $t - 1/2$ will excite a pair of defects $(f_1, t) + (f_2, t)$, where f_1 and f_2 share e. A measurement error on the ancilla qubit $\tilde{e} \in \tilde{E}_t$ at time t will excite a set of defects $\tilde{\partial}_t^\mathsf{T}\tilde{e} \times \{t, t+1\}$. If $\tilde{e}$ is a split edge, $\tilde{\partial}_t^\mathsf{T}\tilde{e}$ consists of only one face, and the error creates a pair of defects correlating time t and $t+1$; otherwise, $\tilde{\partial}_t^\mathsf{T}\tilde{e}$ consists of two faces and four defects are created. We say that the measurement error is of type I if $|\tilde{\partial}_t^\mathsf{T}\tilde{e}| = 1$ or type II if $|\tilde{\partial}_t^\mathsf{T}\tilde{e}| = 2$.

If we always use the Shor-style extraction gadget, we can view each single error as an edge connecting the pair of excitations. The syndrome can be decoded by applying the minimum-weight-perfect-matching (MWPM) algorithm on the decoder graph [37]. In general, however, the existence of type-II errors complicates the decoding problem, as each of them creates four defects instead of two. Note that a type-II error on the ancilla qubit $\tilde{e}$ at time t is equivalent as two data errors on the same data qubit $\gamma_t\tilde{e}$ but at two different times t and $t+1$. We can run MWPM on the decoder graph that consists of data errors and type-I measurement errors, with the cost that each type-II error is regarded as a weight-2 error. Noticing that two different data errors on the same qubit are never on the same logical error loop with minimum weight, treating every type-II error as two data errors will still preserve the code distance:

Theorem 2.—For any choices of transversal gadgets, if there are less than $L/2$ errors in total, MWPM will not introduce a logical error.

The formal proof of Theorem 2 is left in Ref. [36].

In practice, we need to correct the errors after a finite number of measurement rounds. As the information provided by the latest syndrome bits is always unreliable, they will be processed only when further syndrome bits are gathered, which is referred to as the overlapping recovery method [21]. More concretely, we can divide the time axis by some chosen integer time

$$1 = t_1 < t_2 < \cdots < t_i < \cdots.$$

In the ith round of error correction, we decode the syndrome bits in the time interval $[t_i, t_{i+2})$, but only correct the errors in $[t_i - 1/2, t_{i+1} - 1/2)$; then discard the syndrome bits in $[t_i, t_{i+1})$ while the updated syndrome bits in $[t_{i+1}, t_{i+2})$ are kept for the next round of correction. The decoder graph for the ith recovery only contains the vertices (syndrome bits) and edges (errors) in $[t_i - 1/2, t_{i+2} - 1/2)$. The defects can be fused with the time boundary so that its lifetime is extended to the next round. If the distance from time slice t_i to t_{i+1} on the decoder graph, denoted by $d(t_i, t_{i+1})$, is much smaller than the code distance L, it would be too easy for the decoder to fuse a defect at time t_i to the time boundary. Equivalently, the decoder will consider most of the syndrome bit flips as measurement errors so that the data errors can hardly be corrected. Our goal is to correct the data errors in $[t_i - 1/2, t_{i+1} - 1/2)$ before the time t_{i+2}. If $d(t_{i+1}, d_{i+2}) \gg L$, when we have less than $L/2$ errors, the decoder will not fuse a defect before t_{i+1} with another defect after t_{i+2}. Indeed, we have the following result, whose formal proof is in Ref. [36]:

Theorem 3.—For any choices of transversal gadgets and set of integer time $\{t_i\}$, MWPM is fault tolerant only if $d(t_i, t_{i+1}) = \Omega(L)$ for every i.

Our argument generalizes the overlapping recovery method discussed in Ref. [21], which uses Shor error correction for fault tolerance. For Shor error correction, we

have $d(t_i, t_{i+1}) = |t_{i+1} - t_i|$ so that we need to repeat the measurement for $\Omega(L)$ rounds. On the other hand, as an extreme case, if we apply Steane error correction, as there are no type-I errors, the decoder graph does not have any timelike edges. The time slices are not connected with each other so that $d(t, t') = \infty$ for any $t \neq t'$. This allows us to choose $t_i = i$. Moreover, when the MWPM decoder matches the defects in time slice t, the syndrome information later than t is not used at all. Our analysis is compatible with the fact that Steane's scheme supports single-shot error correction [13].

Without taking the complexity of ancilla-state preparation into account, the Steane-style gadget is the best among all transversal gadgets. In practice, however, we might hope that the ancilla block size is bounded by some constant or some function of L. For example, we would like to partition the toric lattice into $m \times m$ square lattices for some m (for simplicity, we assume that m is a divisor of L) so that each ancilla block has $2m(m+1) = O(m^2)$ qubits. Note that the special case $m = 1$ corresponds to the Shor-style gadget. For general m, we naturally expect that the time overhead of fault-tolerant error correction, characterized by $|t_{i+1} - t_i|$, is somewhere between $O(L)$ and $O(1)$. Unfortunately, if the partition is identical for each time slice, as depicted in Fig. 1(a), we will still have $d(t, t') = |t - t'|$ due to the existence of a path of timelike errors $\{(\tilde{e}, t)\}_{t \in \mathbb{N}}$, where $\tilde{e} \in \tilde{\partial} F$ is some split edge. To avoid creating a straight line parallel to the time axis, we can shuffle the type-I errors by shifting the partitions periodically. For simplicity, we assume that $m = 3k$ for some $k \in \mathbb{N}$. We label the rows and columns of the toric lattice from 0 to $L-1$. The vertex on row i and column j is labeled by (i, j). At time $t \in \mathbb{N}$, we obtain a gadget by partitioning the lattice into square lattices of size $m \times m$, whose top-left corners are $((pm + kt) \bmod L, (qm + kt) \bmod L)$, where $p, q = 0, \ldots, L/m - 1$. The decoder graph is demonstrated in Fig. 1(b). One can verify that $d(t, t+3) = t + 2 = \Omega(m)$ for any $t \in \mathbb{N}$. Therefore, it suffices to choose $|t_{i+1} - t_i| = O(L/m)$ to achieve fault tolerance.

We study the circuit-level performance of our fault-tolerant error-correction schemes numerically. The X- and Z-syndrome extractions are applied alternatively. Our error model is parametrized by a single error parameter p and consists of three parts: (1) gate errors, with probability p, each 2-qubit CNOT gate is followed by a Pauli error drawn uniformly at random from the set $\{I, X, Y, Z\}^{\otimes 2} \backslash \{I \otimes I\}$; (2) measurement errors, with probability $2p/3$, a measurement outcome in either the Z or X basis is flipped; (3) preparation errors, ancilla preparation can lead to correlated errors that need to be removed through verification or syndrome measurement decoding. Here we assume a simple error model where the complicated ancilla blocks are generated perfectly and then each qubit undergoes an independent depolarizing channel with probability p.

(a) Aligned ancilla blocks (b) Offset ancilla blocks

FIG. 1. The decoder graphs of the toric code: the syndrome bits are vertices, the data errors are horizontal edges, and the type-I measurement errors are vertical edges. The ancilla blocks when aligned (a) lead to timelike correlations between directed repeated measurements. By offsetting the ancilla blocks (b), the timelike correlations require spacelike errors in order to correlate defects from top to bottom.

Our calculations ignore idling errors, which enables us to avoid complications due to scheduling. Comparison of these syndrome extraction methods for practical application would require a detailed multiparameter error model, a procedure for ancilla generation and verification, and the connectivity constraints of the quantum processor. To accelerate our simulation, instead of the standard MWPM decoding algorithm, we use a weighted variant of the union-find decoder [38,39]. Table II compares the threshold of our block extraction schemes and the conventional Shor's and Steane's schemes. For block extraction, we fix the ancilla block size m when $L \to \infty$. In this case, we will still need $O(L)$ rounds of extractions even if we offset the blocks. However, we observe that offsetting the blocks yields different threshold values than aligning them. This makes sense as the two strategies provide different decoder graph symmetries. When m gets larger, the offset version starts to yield higher threshold values. We also calculate the threshold of the conventional bare-ancilla extraction scheme [40] for a comparison. The raw data of the simulation are available in Ref. [36].

In this Letter, we have constructed a family of syndrome extraction gadgets for CSS codes described by decomposition of parity-check matrices. These gadgets allow us to extract stabilizer elements of the same type in parallel. Notably, our gadget family includes both Shor's [10] and Steane's [12] schemes. We applied these gadgets on the toric code and constructed fault-tolerant error-correction schemes. Remarkably, for a toric code with distance L, one can use ancilla blocks with size $O(m^2)$ to achieve fault tolerance in $O(L/m)$ rounds. By numerical simulation, we found that our error-correction schemes could yield higher thresholds under a simplified circuit-level noise model. As the data qubits would be

TABLE II. Threshold comparison between block syndrome extractions and the standard methods.

	Shor		Block extraction				
Method	Cat	Bare	$m = 3$	$m = 6$	$m = 9$	$m = 12$	Steane
Offset (%)	0.57	0.83	0.68	0.89	1.04	1.13	2.05
Aligned (%)			0.74	0.89	0.97	1.04	

kept idling while preparing the ancilla state, we assume that the idling errors are negligibly small, which could be achieved by trapped-ion qubits [41] or other qubits with long coherence time. In practice, the ancilla preparation errors will depend on the preparation protocols, which can be either direct preparation with postselection [40] or state distillation [42]. A more comprehensive analysis is beyond the scope of this Letter.

We note our method of syndrome extraction is quite general and only requires the codes be CSS. We conjecture that our methods will yield better circuit-level thresholds than the Shor-style syndrome extraction methods currently used on large distance codes such as two-dimensional color codes [30,43], while keeping the rejection rates of post-selection reasonable. However, the decoding problem and the time overhead analysis would be more complicated and code specific. We expect that the framework for analyzing and optimizing the time overhead of sequential Shor-style extractions [25,26] will be helpful for time optimizing Steane-like extraction.

We thank Michael Newman and Rui Chao for helpful discussion. The work was sponsored by the NSF EPiQC Expeditions in Computing (1832377), the Office of the Director of National Intelligence–Intelligence Advanced Research Projects Activity through an Army Research Office Contract (No. W911NF-16-1-0082), and the Army Research Office (W911NF-21-1-0005).

*shilin.huang@duke.edu
†ken.brown@duke.edu

[1] P. W. Shor, Phys. Rev. A **52**, R2493 (1995).
[2] D. Aharonov and M. Ben-Or, in *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing* (1997), p. 176, https://doi.org/10.1145/258533.258579.
[3] E. Knill and R. Laflamme, Phys. Rev. A **55**, 900 (1997).
[4] A. Y. Kitaev, Ann. Phys. (Amsterdam) **303**, 2 (2003).
[5] A. R. Calderbank and P. W. Shor, Phys. Rev. A **54**, 1098 (1996).
[6] A. M. Steane, Phys. Rev. Lett. **77**, 793 (1996).
[7] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane, Phys. Rev. Lett. **78**, 405 (1997).
[8] A. R. Calderbank, E. M. Rains, P. M. Shor, and N. J. A. Sloane, IEEE Trans. Inf. Theory **44**, 1369 (1998).
[9] D. E. Gottesman, Dissertation (Ph.D.) Thesis, California Institute of Technology, 1997.
[10] P. W. Shor, in *Proceedings of 37th Conference on Foundations of Computer Science* (1996), p. 56, https://doi.org/10.1109/SFCS.1996.548464.
[11] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **6**, 97 (2006).
[12] A. M. Steane, Phys. Rev. Lett. **78**, 2252 (1997).
[13] A. M. Steane, Phys. Rev. A **68**, 042322 (2003).
[14] E. Knill, Nature (London) **434**, 39 (2005).
[15] R. Chao and B. W. Reichardt, Phys. Rev. Lett. **121**, 050502 (2018).
[16] R. Chao and B. W. Reichardt, npj Quantum Inf. **4**, 42 (2018).
[17] R. Chao and B. W. Reichardt, PRX Quantum **1**, 010302 (2020).
[18] D. P. DiVincenzo and P. W. Shor, Phys. Rev. Lett. **77**, 3260 (1996).
[19] D. P. DiVincenzo and P. Aliferis, Phys. Rev. Lett. **98**, 020501 (2007).
[20] T. J. Yoder and I. H. Kim, Quantum **1**, 2 (2017).
[21] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, J. Math. Phys. (N.Y.) **43**, 4452 (2002).
[22] H. Bombín, Phys. Rev. X **5**, 031043 (2015).
[23] O. Fawzi, A. Grospellier, and A. Leverrier, in *Proceedings of the 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (2018), p. 743, https://doi.org/10.1145/3434163.
[24] E. T. Campbell, Quantum Sci. Technol. **4**, 025006 (2019).
[25] N. Delfosse, B. W. Reichardt, and K. M. Svore, arXiv:2002.05180.
[26] N. Delfosse and B. W. Reichardt, arXiv:2008.05051.
[27] A. G. Fowler, A. M. Stephens, and P. Groszkowski, Phys. Rev. A **80**, 052312 (2009).
[28] M. Li, M. Gutiérrez, S. E. David, A. Hernandez, and K. R. Brown, Phys. Rev. A **96**, 032341 (2017).
[29] M. Li, D. Miller, and K. R. Brown, Phys. Rev. A **98**, 050301(R) (2018).
[30] C. Chamberland and M. E. Beverland, Quantum **2**, 53 (2018).
[31] C. Chamberland, G. Zhu, T. J. Yoder, J. B. Hertzberg, and A. W. Cross, Phys. Rev. X **10**, 011022 (2020).
[32] S. Huang and K. R. Brown, Phys. Rev. A **101**, 042312 (2020).
[33] C. Chamberland and M. E. Beverland, Quantum **2**, 53 (2018).
[34] Y.-C. Zheng, C.-Y. Lai, T. A. Brun, and L.-C. Kwek, Quantum Sci. Technol. **5**, 045007 (2020).
[35] B. W. Reichardt, arXiv:quant-ph/0406025.
[36] S. Huang and K. R. Brown, companion paper, Phys. Rev. A **104**, 022429 (2021).
[37] J. Edmonds, Can. J. Math. **17**, 449 (1965).
[38] N. Delfosse and N. H. Nickerson, arXiv:1709.06218.
[39] S. Huang, M. Newman, and K. R. Brown, Phys. Rev. A **102**, 012419 (2020).
[40] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Phys. Rev. A **86**, 032324 (2012).
[41] Y. Wang, M. Um, J. Zhang, S. An, M. Lyu, J.-N. Zhang, L.-M. Duan, D. Yum, and K. Kim, Nat. Photonics **11**, 646 (2017).
[42] Y.-C. Zheng, C.-Y. Lai, and T. A. Brun, Phys. Rev. A **97**, 032331 (2018).
[43] H. Bombin and M. A. Martin-Delgado, Phys. Rev. Lett. **97**, 180501 (2006).