Constant overhead quantum fault-tolerance with quantum expander codes

Omar Fawzi
Univ Lyon, ENS de Lyon,
CNRS, UCBL, LIP UMR 5668
F-69007 Lyon, France
omarfawzi@ens-lyon.fr

Antoine Grospellier
Inria
Paris, France
antoine.grospellier@inria.fr

Anthony Leverrier
Inria
Paris, France
anthony.leverrier@inria.fr

Abstract—We prove that quantum expander codes can be combined with quantum fault-tolerance techniques to achieve constant overhead: the ratio between the total number of physical qubits required for a quantum computation with faulty hardware and the number of logical qubits involved in the ideal computation is asymptotically constant, and can even be taken arbitrarily close to 1 in the limit of small physical error rate. This improves on the polylogarithmic overhead promised by the standard threshold theorem.

To achieve this, we exploit a framework introduced by Gottesman together with a family of constant rate quantum codes, quantum expander codes. Our main technical contribution is to analyze an efficient decoding algorithm for these codes and prove that it remains robust in the presence of noisy syndrome measurements, a property which is crucial for fault-tolerant circuits. We also establish two additional features of the decoding algorithm that make it attractive for quantum computation: it can be parallelized to run in logarithmic depth, and is single-shot, meaning that it only requires a single round of noisy syndrome measurement.

Keywords—Decoding algorithm; single-shot quantum error correction; expander codes; percolation;

I. INTRODUCTION

Quantum computers are expected to offer significant, sometimes exponential, speedups compared to classical computers. For this reason, building a large, universal computer, is a central objective of modern science. Despite two decades of effort, experimental progress has been somewhat slow and the largest computers available at the moment reach a few tens of physical qubits, still quite far from the numbers necessary to run “interesting” algorithms. A major source of difficulty is the inherent fragility of quantum information: storing a qubit is quite challenging, processing quantum information even more so.

Any physical implementation of a quantum computer is necessarily imperfect because qubits are subject to decoherence and physical gates can only be approximately realized. In order to perform a correct computation on a faulty computer, techniques of fault-tolerant computation must be developed. One of the crowning achievements of the early days of quantum computing is the threshold theorem which states that upon encoding the logical qubits within the appropriate quantum error correcting code, it is possible to perform arbitrary long computations on a faulty quantum computer, provided that the noise level is below some constant threshold value [1]. This solution comes at a cost, however, since the fault-tolerant version of the circuit to be performed is usually larger that the initial version. In particular, a number of ancilla qubits is required and the space overhead, i.e., the ratio between the total number of physical qubits and the number of logical qubits, scales polylogarithmically with the number of gates involved in the original computation. Indeed, the main technique to achieve fault-tolerance is to protect the logical qubits with concatenated codes. In order to guarantee a final failure probability ε for a circuit C acting on k qubits with |C| locations1, O(log log(|C|/ε)) levels of encoding are needed, which translates into a polylog(|C|/ε) space overhead. While this might seem like a reasonably small overhead, this remains rather prohibitive in practice, and more importantly, it raises the question of whether this value is optimal. In this paper, we consider a realistic model for quantum computing where the quantum gates are noisy, but all classical computation is assumed to be fast and error-free. Note that if classical gates are also noisy, then it is known that classical fault-tolerance cannot be obtained with constant overhead [2], [3].

In a breakthrough paper, Gottesman has shown that the polylogarithmic overhead was maybe not necessary after all, and that polynomial-time computations could be performed with a noisy circuit with only a constant overhead [4]. In fact, the constant can even be taken arbitrarily close to 1 provided that the physical error is sufficiently small. In order to overcome the polylogarithmic barrier, Gottesman suggested to use quantum error correcting codes with constant rate. More precisely, the idea is to encode the logical qubits in large blocks, but still of size sub-linear in k. The encoding can still be made fault-tolerant thanks to concatenation, but this only yields an overhead polylogarithmic in the block size, and choosing a sufficiently small block size to a sub-linear overhead for encoding. Gates are then performed with Knill’s technique by teleporting the appropriate encoded states. Overall, apart from the ini-

1A location refers either to a quantum gate, the preparation of a qubit in a given state, a qubit measurement or a wait location if the qubit is not acted upon at a given time step.
that outputs a distribution as input and yields a quantum code in linear time. More recently in [14], we studied physical qubits, depth \( \varepsilon > 0 \) and a minimum distance equal to that of the classical code or any \( \varepsilon = o(\sqrt{N}) \). Let \( f(k) \) locations for \( f \) an arbitrary polynomial. There exists a circuit \( \tilde{C} \) using \( \eta k \) physical qubits, depth \( O(f(k)) \) and number of locations \( O(kf(k)) \) that outputs a distribution which has total variation distance at most \( \varepsilon \) from the output distribution of \( C \), even if the components of \( \tilde{C} \) are noisy with an error rate \( p < p_T \).

Before moving to the proof techniques, let us mention some limitations of the present work. For our analysis to apply, we need bipartite expander graphs with a large (vertex) expansion. A first issue is that there is no known efficient algorithm that can deterministically construct such graphs\(^2\). While random graphs will display the right expansion (provided their degree is large enough) with high probability, it is not known how to check efficiently that a given graph is indeed sufficiently expanding. The second

\(^2\)While algorithms to construct graphs with large spectral expansion are known, they do not imply a sufficient vertex expansion for our purpose.
issue is that we need graphs with a large (constant) degree, which will translate into significantly large quantum codes. In other words, one shouldn’t expect the present analysis to be applicable to small size quantum codes that might be built in the near future. We note that Gottesman’s analysis also required the initial circuit size to be large enough: this is necessary in order to make the contribution of additive terms sub-linear and therefore obtain a constant overhead. Another limitation of our work is the very small threshold value that it yields. While the threshold is usually expected to lie between $10^{-3}$ and $10^{-2}$ for the best constructions based on code concatenation, we expect our value to be several orders of magnitude smaller, as this was already the case in Gottesman’s paper [4] and in our previous work with perfect syndrome measurement [14]. Part of the explanation is due to the very crude bounds that we obtain via percolation theory arguments. In this work, we haven’t tried to optimize the value of the threshold and have instead tried to simplify the general scheme as much as possible. We expect that future work, in particular based on simulations, will help to better understand the true value of the threshold for fault-tolerance schemes with constant overhead. Finally, as already pointed out, we consider a model with error-free classical computation, and assume that the logarithmic-depth decoding algorithm can be performed in constant time.

Main result and proof techniques

In this section, we provide an informal overview of the main result and the techniques used for the proofs. More formal definitions and proofs can be found in the following sections. When decoding a quantum error correcting code, two types of errors need to be taken into account: the $X$-type (or bit flip) errors and the $Z$-type (or phase flip) errors. They play a symmetric role for the codes we consider and it is therefore sufficient to focus on bit flips for instance. An $X$-type error is described by a subset $E$ of the qubits to which the bit flip operator $X$ was applied. To decode a quantum error correcting code, we start by performing a measurement that returns the syndrome $\sigma = \sigma(E)$ which only depends on the error. The objective of the decoding algorithm is given $\sigma$ to output an error $\hat{E}$ which is the same as $E$. More precisely, it is sufficient for the errors $E$ and $\hat{E}$ to be equivalent in the sense that the error $E \oplus \hat{E}$ acts trivially on every codeword (for a stabilizer code, this simply means that $E \oplus \hat{E}$ belongs to the stabilizer group). As previously mentioned, our main contribution is to analyze the small-set-flip decoding algorithm in the setting where the syndrome measurement is noisy, i.e., the decoding algorithm takes as input $(\sigma(E) \oplus D)$ instead of just $\sigma(E)$, where $D$ represents the syndrome measurement error. The objective of the decoding algorithm is then not to recover the error exactly (which will not be possible) but rather to control the size of remaining error $E \oplus \hat{E}$. In the context of quantum fault-tolerance, the relevant error model for the pair $(E, D)$ is the local stochastic noise model with parameters $(p, q)$ defined by requiring that for any $F$ and $G$, the probability that $F$ and $G$ are part of the qubit and syndrome errors, respectively, is bounded as follows, $\mathbb{P}[F \subseteq E, G \subseteq D] \leq p|F|q|G|$

**Theorem 2 (Informal).** There exist constants $p_0 > 0, p_1 > 0$ such that the following holds. Consider a bipartite graph with sufficiently good expansion and the corresponding quantum expander code. Consider random errors $(E, D)$ satisfying a local stochastic noise model with parameter $(p_{\text{phys}}, p_{\text{synd}})$ with $p_{\text{phys}} < p_0$ and $p_{\text{synd}} < p_1$. Let $\hat{E}$ be the output of the small-set-flip decoding algorithm on the observed syndrome. Then, except for a failure probability of $e^{-\Omega(\sqrt{n})}$, $E \oplus \hat{E}$ is equivalent to $E_\text{err}$ that has a local stochastic distribution with parameter $p_{\text{synd}}$. In addition, the small-set-flip algorithm can be parallelized to run in $O((\log n))$ depth.

In the special case where the syndrome measurements are perfect, i.e., $p_{\text{synd}} = 0$, the statement guarantees that for a typical error of size at most $p_0 n$, the small-set-flip algorithm finds an error that is equivalent to the error that occurred. If the syndrome measurements are noisy, then we cannot hope to recover an equivalent error exactly, but instead we can control the size of the remaining error $E \oplus \hat{E}$ by the amount of noise in the syndrome measurements. In particular, for any qubit error rate below $p_0$, the decoding operation reduces this error rate to be $p_{\text{synd}}$ (our choice of $p_0$ will be such that $p_{\text{synd}} \leq p_0$). This criterion is sufficient for fault-tolerant schemes as it ensures that the size of the qubit errors stay bounded throughout the execution of the circuit. The proof of this theorem consists of two main parts: analyzing arbitrary low weight errors below the minimum distance (Proposition 10) and exploiting percolation theory to analyze stochastic errors of linear weight (Theorem 9).

The small-set-flip decoding algorithm proceeds by trying to flip small sets of qubits so as to decrease the weight of the syndrome, and the main challenge in its analysis is to prove the existence of such a small set $F$. In the case where the observed syndrome is error free, Refs [13] and [14] relied on the existence of a “critical generator” to exhibit such a set of qubits. This approach, however, only yields a single such set $F$, and when the syndrome becomes noisy, nothing guarantees anymore that flipping the qubits in $F$ will result in a decrease of the syndrome weight and it becomes unclear whether the decoding algorithm can continue. Instead, in order to take into account the errors on the syndrome measurements, we would like to show that there are many possible sets of qubits $F$ that decrease the syndrome weight. In order to establish this point, we consider an error $E$ of size below the minimum distance and we imagine running the (sequential) decoding algorithm [13] without errors on the syndrome. The algorithm gives a sequence of small sets
to flip successively in order to correct the error. In other words, we obtain the following decomposition of the error, \( E = \bigoplus_i F_i \) (note that the sets \( F_i \) might overlap). The expansion properties of the graph guarantee that there are very few intersections between the syndromes \( \sigma(F_i) \) (see the proof of Proposition 10). In particular in the case of noiseless syndrome, a linear number of these \( F_i \) can be flipped to decrease the syndrome weight. There are two consequences to this result. First, it is possible to parallelize the decoding algorithm by flipping multiple \( F_i \) in each round and this decreases the syndrome weight by a constant factor, thereby correcting the error after a logarithmic number of rounds. Second, even when the syndrome is noisy there will remain some \( F_i \) that can be flipped in order to decrease the syndrome weight and finally, the size of the error \( E \oplus E \) can be upper bounded with a linear function of the syndrome error size (Proposition 10).

In order to analyze random errors of linear weight, we show using percolation theory that, with high probability, the error forms clusters in the sense of connected \( \alpha \)-subsets (Definition 16 and Lemma 23). This is similar to the analysis in [16], [14], except that we use the syndrome adjacency graph of the code (as in [4]) to establish the “locality” of the decoding algorithm, implying that each cluster of the error is corrected independently of the other ones (Lemma 15). Using the fact that clusters are of size bounded by the minimum distance of the code, the result on low weight errors shows that the size of \( E \oplus E \) is controlled by the syndrome error size. In order to show that the error after correction is local stochastic, we introduce the notion of witness (Definition 20). The basic idea is to find a syndrome error in the neighborhood of a given qubit error \( S \). However, a qubit error can be the consequence of a distant syndrome error. This is why a witness is defined as a set of qubit errors \( W \) but is potentially larger than \( S \). The previously mentioned results show that witnesses exist for \( E \oplus E \) and we conclude our proof using an upper bound on the probability that a witness exists.

A remarkable feature of our analysis is that it shows that the small-set-flip decoding algorithm only uses a single noisy syndrome measurement and outputs an error with controlled weight. Note that this is in contrast to decoding algorithms for many other codes such as the toric code for which such a repetition is necessary. This property is called single-shot in the fault-tolerant quantum computation literature [15], [17].

**Organization**

We start in Section II with notations and preliminaries to recall the construction and main properties of quantum expander codes, their efficient decoding algorithm. We also introduce the relevant noise model. In Section III, we establish our main technical result showing that the small-set-flip decoding algorithm for quantum expander codes is robust against local stochastic noise in the syndrome measurement. Due to space limitations, some of the proofs as well as the parallelization of the decoding algorithm are deferred to the full version of the paper [18].

**II. Preliminaries**

In this section, we first review the construction of classical and quantum expander codes. We then discuss models of noise which are relevant in the context of quantum fault-tolerance.

A. Classical expander codes

A linear classical error correcting code \( C \) of dimension \( k \) and length \( n \) is a subspace of \( \mathbb{F}_2^n \) of dimension \( k \). Mathematically, it can be defined as the kernel of an \( (n - k) \times n \) matrix \( H \), called the parity-check matrix of the code: \( C = \{ x \in \mathbb{F}_2^n : Hx = 0 \} \). The minimum distance \( d_{min} \) of the code is the minimum Hamming weight of a nonzero codeword: \( d_{min} = \min \{ |x| : x \in C, x \neq 0 \} \). Such a linear code is often denoted as \( [n, k, d_{min}] \). It is natural to consider families of codes, instead of single instances, and study the dependence between the parameters \( n, k \) and \( d_{min} \). In particular, a family of codes has constant rate if \( k = \Theta(n) \). Another property of interest of a linear code is the weight of the rows and columns of the parity-check matrix \( H \). If these weights are upper bounded by a constant, then we say that the code is a low-density parity-check (LDPC) code [19]. This property is particularly attractive because it allows for efficient decoding algorithms, based on message passing for instance.

An alternative description of a linear code is via a bipartite graph known as its factor graph. Let \( G = (V \cup C, E) \) be a bipartite graph, with \( |V| = n_V \) and \( |C| = n_C \). With such a graph, we associate the \( n_C \times n_V \) matrix \( H \), whose rows are indexed by the vertices of \( C \), whose columns are indexed by the vertices of \( V \), and such that \( H_{vw} = 1 \) if \( v \) and \( w \) are adjacent in \( G \) and \( H_{vw} = 0 \) otherwise. The binary linear code \( C_G \) associated with \( G \) is the code with parity-check matrix \( H \). The graph \( G \) is the factor graph of the code \( C_G \). \( V \) is the set of bits and \( C \) is the set of check-nodes.

It will sometimes be convenient to describe codewords and error patterns as subsets of \( V \): the binary word \( e \in \mathbb{F}_2^n \) is described by a subset \( E \subseteq V \) whose indicator vector is \( e \). Similarly we define the syndrome of a binary word either as a binary vector of length \( n_C \) or as a subset of \( C \):\n
\[
\sigma(e) := He \in \mathbb{F}_2^{n_C}, \quad \sigma(E) := \bigoplus_{v \in E} \Gamma(v) \subseteq C,
\]

where \( \Gamma(v) \subseteq C \) is the set of neighbors of \( v \). In this paper, the operator \( \bigoplus \) is interpreted either as the symmetric difference of sets or as the bit-wise exclusive disjunction depending on whether errors and syndromes are interpreted as sets or as binary vectors.
A family of codes that will be central in this work are those associated to so-called expander graphs, that were first considered by Sipser and Spielman in [12].

**Definition 3 (Expander graph).** Let \( G = (V \cup C, \mathcal{E}) \) be a bipartite graph with left and right degrees bounded by \( d_L \) and \( d_C \) respectively. Let \( |V| = nV \) and \( |C| = nC \). We say that \( G \) is \((\gamma, \delta)\)-left-expanding for some constants \( \gamma, \delta > 0 \), if for any subset \( S \subseteq V \) with \( |S| \leq \gamma nV \), the neighborhood \( \Gamma(S) \) of \( S \) in the graph \( G \) satisfies \([\Gamma(S)] \geq (1 - \delta)d_V|S|\). Similarly, we say that \( G \) is \((\gamma, \delta)\)-right-expanding if for any subset \( S \subseteq C \) with \( |S| \leq \gamma nC \), we have \([\Gamma(S)] \geq (1 - \delta)d_C|S|\). Finally, the graph \( G \) is said \((\gamma, \delta)\)-expanding if it is both \((\gamma, \delta)\)-left expanding and \((\gamma, \delta)\)-right expanding.

Sipser and Spielman introduced expander codes, which are the linear codes associated with (left-)expander graphs. Remarkably these codes come with an efficient decoding algorithm that can correct arbitrary errors of weight \( \Omega(n) \) [12].

**B. Quantum error correcting codes**

A quantum code encoding \( k \) logical qubits into \( n \) physical qubits is a subspace of \((\mathbb{C}^2)^\otimes n\) of dimension \( 2^k \). A quantum stabilizer code is described by a stabilizer, that is an Abelian group of \( n\)-qubit Pauli operators (tensor products of single-qubit Pauli operators \( X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \) and \( Y = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \)) and \( I \) with an overall phase of \( \pm 1 \) or \( \pm i \) that does not contain \(-I\). The code is defined as the eigenspace of the stabilizer with eigenvalue \(+1\) [20]. A stabilizer code of dimension \( k \) can be described by a set of \( n-k \) generators of its stabilizer group.

A particularly nice construction of stabilizer codes is given by the CSS construction [21], [22], where the stabilizer generators are either products of single-qubit \( X \)-Pauli matrices or products of \( Z \)-Pauli matrices. The condition that the stabilizer group is Abelian therefore only needs to be enforced between \( X \)-type generators (corresponding to products of Pauli \( X \)-operators) and \( Z \)-type generators. More precisely, consider two classical linear codes \( C_X \) and \( C_Z \) of length \( n \) satisfying \( C_Z^\perp \subseteq C_X \) (or equivalently, \( C_X^\perp \subseteq C_Z \)). The dual code \( C_X^\perp \) to \( C_X \) consists of the words which are orthogonal to all the words in \( C_X \). This condition also reads \( H_X \cdot H_Z^\perp = 0 \), if \( H_X \) and \( H_Z \) denote the respective parity-check matrices of \( C_X \) and \( C_Z \). The quantum code \( CSS(C_X,C_Z) \) associated with \( C_X \) (used to correct \( X \)-type errors and corresponding to \( Z \)-type stabilizer generators) and \( C_Z \) (used to correct \( Z \)-type errors and corresponding to \( X \)-type stabilizer generators) has length \( n \) and is defined as the linear span of \( \left\{ \sum_{x \in C_X} \left[ x + z \right] : x \in C_X \right\} \), where \( \{x\} : x \in \mathbb{F}_2^n \) is the canonical basis of \((\mathbb{C}^2)^\otimes n\). In particular, two states differing by an element of the stabilizer group are equivalent. The dimension of the CSS code is given by \( k = \dim(C_X/C_Z^\perp) = \dim(C_Z/C_X^\perp) = \dim(C_X + \dim(C_Z - n) \). Its minimum distance is defined in analogy with the classical case as the minimum number of single-qubit Pauli operators needed to map a codeword to an orthogonal one. For the code \( CSS(C_X,C_Z) \), one has \( d_{\min} = \min\{d_X, d_Z\} \) where \( d_X = \min\{|E| : E \in C_X \setminus C_Z^\perp\} \) and \( d_Z = \min\{|E| : E \in C_Z \setminus C_X^\perp\} \). We say that \( CSS(C_X,C_Z) \) is a \([n, k, d_{\min}] \) quantum code. In the following, it will be convenient to consider the factor graph \( G_X = (V \cup C_X, \mathcal{E}_X) \) (resp. \( G_Z \)) of \( C_X \) (resp. of \( C_Z \)). We will denote by \( \Gamma_X \) (resp. \( \Gamma_Z \)) the neighborhood in \( G_X \) (resp. \( G_Z \)). For instance, if \( g \in C_Z \) is an \( X \)-type generator, that is a product of Pauli \( X \) operators, then \( \Gamma_Z(g) \) is the set of qubits (indexed by \( V \)) on which the generator acts non-trivially.

Among stabilizer codes, and CSS codes, the class of quantum LDPC codes stands out for practical reasons: these are the codes for which one can find sparse parity-check matrices \( H_X \) and \( H_Z \). More precisely, such matrices are assumed to have constant row weight and constant column weight.

A natural noise model is the so-called Pauli-type noise, mapping a qubit \( \rho \) to \( p_X X \rho + p_Y Y \rho + p_Z Z \rho Z \rho \) for some \( p_X, p_Y, p_Z \). Such a noise model is particularly convenient since one can interpret the action of the noise as applying a given Pauli error with some probability. As usual, it is sufficient to deal with both \( X \) and \( Z \)-type errors in order to correct Pauli-type errors, and one can therefore define an error by the locations of the Pauli \( X \) and Pauli \( Z \) errors. An error pattern is a pair \((E_X, E_Z)\) of \( n \)-bit strings which describe the locations of the Pauli \( X \) errors, and Pauli \( Z \) errors respectively. The syndrome associated with \((E_X, E_Z)\) for the code \( CSS(C_X, C_Z) \) consists of \( \sigma_X = \sigma_X(E_X) := H_X E_X \) and \( \sigma_Z = \sigma_Z(E_Z) := H_Z E_Z \). A decoder is given the pair \((\sigma_X, \sigma_Z)\) of syndromes and should return a pair of errors \((\hat{E}_X, \hat{E}_Z)\) such that \( E_X + \hat{E}_X \in C_Z^\perp \) and \( E_Z + \hat{E}_Z \in C_X^\perp \). In that case, the decoder outputs an error equivalent to \((E_X, E_Z)\), and we say that it succeeds. Similarly as in the classical case, it will be convenient to describe \( X \)-type error patterns and \( X \)-type syndromes as subsets of the vertices of the factor graph \( G_X = (V \cup C_X, \mathcal{E}_X) \). The error pattern is then described by a subset \( E_X \subseteq V \) whose syndrome is the subset \( \sigma_X(E_X) \subseteq C_X \) defined by \( \sigma_X(E_X) := \bigoplus_{v \in E_X} \Gamma_X(v) \). One describes \( Z \)-type error patterns and \( Z \)-type syndromes in the same fashion using the factor graph \( G_Z \).

In this paper, we consider Algorithm 2 which tries to recover \( E_X \) and \( E_Z \) independently. More precisely, the algorithm is given by an \( X \)-decoding algorithm that takes as input \( \sigma_X \) and returns \( \hat{E}_X \) such that \( \sigma_X(\hat{E}_X) = \sigma_X \), and a \( Z \)-decoding algorithm that takes as input \( \sigma_Z \) and returns \( \hat{E}_Z \) such that \( \sigma_Z(\hat{E}_Z) = \sigma_Z \). Here the two algorithms are identical upon exchanging the roles of \( X \) and \( Z \). We note that this kind of decoding algorithm might achieve sub-optimal error probabilities for some error models. In fact, if there are correlations between \( X \) and \( Z \) errors (for instance in the case of the depolarizing channel where
\(p_X = p_Y = p_Z\), one can decrease the error probability by trying to recover \(E_X\) by using both \(\sigma_X\) and \(\sigma_Z\).

### C. Quantum expander codes

In this work, we are particularly interested in a family of LDPC CSS codes that features a constant rate and a minimum distance \(\Theta(\sqrt{n})\) obtained by applying the hypergraph product construction of Tillich and Zémor to classical expander codes. If these expander codes have sufficient expansion, the corresponding quantum code is called quantum expander code and comes with an efficient decoding algorithm [13].

The construction is as follows. Let \(G = (A \cup B, E)\) be a biregular \((\gamma, \delta)\)-expanding graph with \(\delta\) sufficiently small \(^3\), and constant left and right degrees denoted \(d_A\) and \(d_B\) with \(d_A \leq d_B\). Let us also denote \(n_A = |A|\) and \(n_B = |B|\) with \(n_B \leq n_A\). Such graphs can be found in a probabilistic fashion provided that \(d_A \geq [\delta^{-1}]\). Let \(C\) be the classical code associated with \(G\), let \(d_{\text{min}}(C)\) be the minimal distance of \(C\) and let \(H\) be its parity-check matrix (that we assume to be full rank) corresponding to the factor graph \(G\). In particular, the weights of rows and columns of \(H\) are \(d_A\) and \(d_B\), respectively. The hypergraph product code of \(C\) with itself admits the following parity check matrices:

\[
H_X = (I_{n_A} \otimes H, H^T \otimes I_{n_B}),
\]

\[
H_Z = (H \otimes I_{n_A}, I_{n_B} \otimes H^T).
\]

The code is LDPC with generators of weight \(d_A + d_B\) and qubits involved in at most \(2d_B\) generators.

The following theorem summarizes the main properties of this quantum code.

**Theorem 4** (Tillich, Zémor [7]). The CSS code defined above is LDPC with parameters \([n, k, d_{\text{min}}]\), where \(n = n_A^2 + n_B^2\), \(k = (n_B - n_B)\), and \(d_{\text{min}} = d_{\text{min}}(C)\).

A natural approach to perform error correction would be to directly mimic the classical bit-flip decoding algorithm analyzed by Sipser and Spielman, that is to try applying \(X\)-type (or \(Z\)-type) correction to qubits when it leads to a decrease of the syndrome weight. Unfortunately, in that case, there are error configurations of constant weight that couldn’t be corrected. This suggests the “small-set-flip” strategy that we describe next.

Focusing on \(X\)-type errors for instance, and assuming that the syndrome \(\sigma = H_X E\) is known, the algorithm cycles through all the \(X\)-type generators of the stabilizer group (i.e. the rows of \(H_Z\)), and for each one of them, determines whether there is an error pattern contained in the generator that decreases the syndrome weight. Assuming that this is the case, the algorithm applies the error pattern (choosing the one maximizing the ratio between the syndrome weight decrease and the pattern weight), if there are several. The algorithm then proceeds by examining the next generator.

Since the generators have (constant) weight \(d_A + d_B\), there are \(2^{d_A + d_B}\) possible patterns to examine for each generator.

Before describing the algorithm more precisely, let us introduce some additional notations. Let \(X\) be the set of subsets of \(V\) corresponding to \(X\)-type generators: \(X = \{T_X(g) : g \in C_Z\} \subseteq P(V)\), where \(P(V)\) is the power set of \(V\). The indicator vectors of the elements of \(X\) span the dual code \(C_Z^\perp\). The condition for successful decoding of the \(X\) error \(E\) then asks that there exists a subset \(X \subseteq X\) such that \(E \oplus \hat{E} = \bigoplus_{x \in X} x\), where \(\hat{E}\) is the output of the decoding algorithm. This means that the remaining error after decoding is trivial, that is equal to a sum of generators. At each step, the small-set-flip algorithm tries to flip a subset of \(T_X(g)\) for some generator \(g \in C_Z\) which decreases the syndrome weight \(|\sigma|\). In other words, it tries to flip some qubit \(F \in F_0\) such that \(\Delta(\sigma, F) > 0\) where:

\[F_0 := \{F \subseteq T_X(g) : g \in C_Z\}, \]

\[\Delta(\sigma, F) := |\sigma| - |\sigma + \sigma_X(F)|. \]

The decoding algorithm consists of two iterations of Algorithm 1 below: it first tries to correct \(X\)-type errors, then it is applied a second time (exchanging the roles of \(X\) and \(Z\)) to correct \(Z\)-type errors.

**Algorithm 1** (Ref. [13]): Small-set-flipping decoding algorithm for quantum expander codes

**INPUT:** \(\sigma \subseteq C_X\), a syndrome where \(\sigma = \sigma_X(E)\) with \(E \subseteq V\) an error

**OUTPUT:** \(\hat{E} \subseteq V\), a guess for the error pattern (alternatively, a set of qubits to correct)

**SUCCESS:** if \(E \oplus \hat{E} = \bigoplus_{x \in X} x\) for \(X \subseteq X\), i.e. \(E\) and \(\hat{E}\) are equivalent errors

\[E_0 = \emptyset ; \sigma_0 = \sigma ; i = 0\]

while \((\exists F \in F_0 : \Delta(\sigma_i, F) > 0)\) do

\[F_i = \arg \max_{F \in F_0} \frac{\Delta(\sigma_i, F)}{|F|}\]

\[\hat{E}_{i+1} = \hat{E}_i + F_i\]

\[\sigma_{i+1} = \sigma_i + \sigma_X(F_i)\]

\[i = i + 1\]

**end while**

**return** \(\hat{E}_i\)

It was proven in Ref. [13] that this algorithm corrects arbitrary errors of size \(O(\sqrt{n})\) provided that the expansion of the graph satisfies \(\delta < 1/6\).

**Theorem 5** (Leverrier, Tillich, Zémor [13]). Let \(G = (A \cup B, E)\) be a \((d_A, d_B)\)-biregular \((\gamma, \delta)\)-expanding graph

\[p_X = p_Y = p_Z\], one can decrease the error probability by trying to recover \(E_X\) by using both \(\sigma_X\) and \(\sigma_Z\).
with $\delta < 1/6$. Letting $d_A$ and $d_B$ be fixed, the small-set-flip decoding algorithm (Algorithm 1) runs in time linear in the code length $n = n_A^2 + n_B^2$, and decodes any quantum error pattern of weight less than $w_0 = \frac{n_A n_B}{(n_A + n_B)^2}$.

In a recent work, the analysis was extended to the case of random errors (either independent and identically distributed, or local stochastic) provided that the syndrome extraction is performed perfectly and under a stricter condition on the expansion of the graph [14].

**Theorem 6** (Fawzi, Grosepellier, Leverrier [14]). Let $G = (A \cup B, E)$ be a $(d_A, d_B)$-biregular $(\gamma, \delta)$-expanding graph with $\delta < 1/8$. Then there exists a probability $p_0 > 0$ and constants $C, C'$ such that if the noise parameter on the qubits satisfies $p < p_0$, the small-set-flip decoding algorithm (Algorithm 1) runs in time linear in the code length and corrects a random error with probability at least

$$1 - C n \left( \frac{p}{p_0} \right)^{C' \sqrt{n}}.$$

The caveat of this result is that it only applies in absence of errors for the syndrome extraction. The main technical contributions of this paper are to establish that the same algorithm still works in presence of noise on the syndrome, and to show that the decoding algorithm can be parallelized to run in logarithmic time.

**D. Noise models**

In the context of quantum fault-tolerance, we are interested in modeling noise occurring during a quantum computation. We refer the reader to the introduction on the topic by Gottesman for a thorough description of noise models for fault-tolerance [23]. In the circuit model of quantum computation, the effect of noise is to cause faults occurring at different locations of the circuit: either on the initial state and ancillas, on gates (either active gates or storage gates) or on measurement gates. We refer to this model as basic model for fault-tolerance. The main idea to perform a computation in a fault-tolerant manner is then to encode the logical qubits with a quantum error correcting code, replace the locations of the original circuit by gadgets applying the corresponding gate on the encoded qubits, and interleave the steps of the computation with error correction steps. In general, it is convenient to abstract away the details of the implementation and consider a simplified model of fault-tolerance where one is concerned with only two types of errors: errors occurring at each time step on the physical qubits, and errors on the results of the syndrome measurement. The link between the basic and the simplified models for fault-tolerance can be made once a specific choice of gate set and gadgets for each gate is made. This is done for instance in Section 7 of Ref. [4]. In other words, the simplified model of fault-tolerance allows us to work with quantum error correcting codes where both the physical qubits and the check nodes are affected by errors.

As usual in the context of quantum error correction, we restrict our attention to Pauli-type errors since the ability to correct all Pauli errors of weight $t$ implies that arbitrary errors of weight $t$ can be corrected. In particular, one only needs to address $X$ and $Z$-type errors since a $Y$-error corresponds to simultaneous $X$ and $Z$-errors. Therefore, we think of an error pattern on the qubits as a pair $(E_X, E_Z)$ of subsets of the set of qubits $V$. This should be interpreted as Pauli error $X$ on all qubits in $E_X \setminus E_Z$, error $Y$ on $E_X \cap E_Z$ and error $Z$ on $E_Z \setminus E_X$. Similarly, the error on the syndrome consists of two classical strings $(D_X, D_Z)$ which are subsets of the sets $C_X$ and $C_Z$ of check nodes, whose values have been flipped. This means that the syndromes that are provided as the input of the decoding algorithm are

$$\sigma_X := \sigma_X(E_X) \oplus D_X, \quad \sigma_Z := \sigma_Z(E_Z) \oplus D_Z.$$  \hspace{1cm} (2)

The algorithm we will consider in this work treat $X$ errors and $Z$ errors in a symmetric fashion. More precisely, the decoding algorithm first tries to recover $E_X$ from $\sigma_X$, then proceeds in a similar way to try to recover $E_Z$ from $\sigma_Z$, without exploiting any information about $E_X$ or $\sigma_X$. Said otherwise, the algorithm tries to recover both $E_X$ and $E_Z$ independently. For this reason, it will be convenient to restrict our attention to $X$-type errors in the following since $Z$-type error would be treated in the same way. In particular, an error will correspond to two sets: a subset $E$ of the qubits and a subset $D$ of the check nodes.

**Definition 7** (Local stochastic error model). Let $V$ be the set of qubits and $C$ be the set of check nodes. A random error $(E, D)$ with $E \subseteq V$ and $D \subseteq C$ satisfies the local stochastic error model with parameters $(p, q)$ if for all $S \subseteq V$ and $T \subseteq C$, we have

$$P[S \subseteq E, T \subseteq D] \leq p^{|S|}q^{|T|}. \hspace{1cm} (3)$$

If $q = 0$, i.e., there are no errors on the syndrome, then we talk of a local stochastic model of parameter $p$. In other words, the location of the errors is arbitrary but the probability of a given error decays exponentially with its weight.

In this paper, we study a variant of Algorithm 1 that allows us to deal with syndrome errors. This is Algorithm 2 below (see eqs. (1) and (4) for notations). The three differences with Algorithm 1 are the input $\sigma = \sigma_X(E) \oplus D$ instead of $\sigma = \sigma_X(E)$, while loop condition $\Delta(\sigma_i, F_i) \geq \beta |\sigma_X(F_i)|$ instead of $\Delta(\sigma_i, F_i) > 0$ and the use of $F$ instead of $F_0$ as set of possible flips (see Remark 8 for a discussion about these changes):

$$F := \left\{ F \subseteq \Gamma_Z(g) : g \in C_Z \text{ and } |\sigma_X(F)| \geq \frac{d_A}{2} |F| \right\}. \hspace{1cm} (4)$$
Algorithm 2: Small-set-flip decoding algorithm for quantum expander codes of parameter $\beta \in (0; 1]$

INPUT: $\sigma \subseteq C_X$ a syndrome such that $\sigma = \sigma_X(E) \oplus D$ for some (unknown) $E \subseteq V$ and $D \subseteq C_X$

OUTPUT: $\hat{E} \subseteq V$, a guess for the error pattern (alternatively, a set of qubits to correct)

\[
\begin{align*}
E_0 &= 0 \ ; \ s_0 = \sigma \ ; \ i = 0 \\
\text{while} \ \exists F_i \in \mathcal{F} : \Delta(\sigma_i, F_i) \geq \beta|\sigma_X(F_i)| \ \text{do} \\
\ & E_{i+1} = E_i \oplus F_i \\
\ & \sigma_{i+1} = \sigma_i \oplus \sigma_X(F_i) \\
\ & \mathcal{F}_{i+1} = \mathcal{F}_{i+1} \setminus \mathcal{F}_i \\
\ & i = i + 1 \\
\end{align*}
\]

end while

Remark 8. In order to simplify our discussion in the paper, we will say that the input of Algorithm 2 is $(E, D)$ when its input is $\sigma_X(E) \oplus D$, we will call $\hat{E}$ the output, we will call $E \oplus \hat{E}$ the remainder, we denote by $f$ the number of steps and we will call $U = E \cup F_0 \cup \ldots \cup F_{f-1}$ the execution support.

Using $\mathcal{F}$ instead of $\mathcal{F}_0$ in Algorithm 2 is not restrictive because if the condition $|\sigma_X(F)| \geq \frac{d_B}{2\delta}|F|$ is not satisfied for some $F \subseteq \Gamma_Z(g)$ then this condition is satisfied by $F^* = \Gamma_Z(g) \setminus F$ (see the proof of Lemma 14).

In Algorithm 1, the weaker “while loop condition” $\Delta(\sigma_i, F_i) > 0$ was used, but it turns out that if $D = \emptyset$ then with high probability on the choice of $E$, the condition $\Delta(\sigma_i, F_i) \geq (1 - 8\delta)|\sigma_X(F_i)|$ is automatically satisfied at each step of Algorithm 1 (this property was used in the proof of Theorem 6). On the other hand when $D \neq \emptyset$, requiring $\Delta(\sigma_i, F_i) \geq \beta|\sigma_X(F_i)|$ with $\beta$ close to 1 makes Algorithm 2 more robust against syndrome errors.

The behavior of Algorithm 2 in the particular case where $D = \emptyset$ could be studied following the proof of [14]: given $Q_G$, a quantum expander code constructed using $G$ some bipartite $(\gamma, \delta)$-expander graph with $\delta < 1/8$, it is possible to prove that a random error $E$ is corrected with high probability by the small-set-flip algorithm of parameter $\beta_0$ ($\beta_0$ as defined in Section III-A). In the noisy case $D \neq \emptyset$, we cannot hope to entirely correct the error because any single qubit error cannot be distinguished from a well-chosen constant weight syndrome bit error. But we will prove in Theorem 9 that when $\delta < 1/16$, the correction provided by the small-set-flip algorithm of parameter $\beta < \beta_1$ ($\beta_1$ as defined in Section III-A) leads to a residual error that is local stochastic with controlled parameter.

III. ANALYSIS OF ALGORITHM 2

A. Notations

The algorithms of Section III depend on parameters $\delta, \beta \in (0; 1]$

We consider $G = (A \cup B, E)$ a $(d_A, d_B)$-biregular $(\gamma, \delta)$-expander graph with $\gamma > 0$, we denote by $Q$ the quantum expander code associated to $G$ (see Section II-C) and by $V, C_X, C_Z$ and $n := |V|$ respectively the set of qubits, the set of $Z$-type stabilizer generators, the set of $X$-type stabilizer generators and the number of physical qubits of $Q$. We will also use $\Gamma_X$ and $\Gamma_Z$ the neighborhoods in the graphs $G_X$ and $G_Z$ as defined in Section II-C.

We run the small-set-flip decoding algorithm (Algorithm 2) of parameter $\beta$ on input $(E, D)$ where $E \subseteq V$ represents a qubit error and $D \subseteq C_X$ represents a syndrome error, we denote by $\hat{E}$ the output of the algorithm, by $f$ the number of steps and by $U = E \cup F_0 \cup \ldots \cup F_{f-1}$ the execution support.

We also define the constants:

\[
\begin{align*}
r &= \frac{d_A}{d_B}, \\
\gamma_0 &= \frac{r^2}{\sqrt{1 + r^2}}, \\
\beta_0 &= \beta_0(\delta) := 1 - 8\delta, \\
\beta_1 &= \beta_1(\delta) := 1 - 16\delta, \\
c_0 &= c_0(\beta, \delta) := \frac{4}{d_A(\beta_1 - \beta)}, \\
c_1 &= c_1(\beta, \delta) := \frac{\beta_1 - \beta}{\beta_0(1 - \beta)}, \\
\alpha_0 &= \alpha_0(\beta) := \frac{r\beta}{4 + 2r\beta}.
\end{align*}
\]

B. Statements of the theorems

In this section, we are going to prove Theorem 9, a generalized version of Theorem 6 that we can apply in the case where the syndrome error $D \subseteq C_X$ is not empty.

Theorem 9. We use the notations of Section III-A with $\delta < 1/16$ and $\beta < \beta_1$.

There exist constants $p_0 > 0, p_1 > 0$ such that the following holds. Suppose the pair $(E, D)$ satisfies a local stochastic noise model with parameter $(p_{\text{phys}}, p_{\text{synd}})$ where $p_{\text{phys}} < p_0$ and $p_{\text{synd}} < p_1$. Then there exists an event such that has probability $1 - e^{-\Omega(\sqrt{n})}$ and a random variable $E_{\text{phys}}$ that is equivalent to $E \oplus \hat{E}$ such that conditioned on $\text{success,}$, $E_{\text{phys}}$ has a local stochastic distribution with parameter $K \cdot p_{\text{synd}}$ where $K$ is a constant independent of $p_{\text{synd}}$.

C. Small adversarial errors

The first step to prove Theorem 9 is to study the case where the qubit error $E$ can be adversarial but where $E \oplus \hat{E}$ is supposed to be reduced with $|E \oplus \hat{E}| \leq \gamma_0 \sqrt{n}$. Here “reduced” means that $E \oplus \hat{E}$ has the smallest Hamming weight among all errors equivalent to $E$. The result in that case is summarized in Corollary 12: it is possible to use expansion-based arguments to find an upper bound on $|E \oplus \hat{E}|$ which grows linearly with $|E \oplus \hat{E}|$. Corollary 12 is a consequence of Proposition 10 and Lemma 11 that we state now but only prove at the end of this section.

Proposition 10. We use the notations of Section III-A with $\delta < 1/16$ and $\beta \in (0; \beta_1)$.
If \(|E| \leq \gamma_0 \sqrt{n}\) and \(|\sigma_X(E)| > c_2 |D \cap \sigma_X(E)|\) then there exists at least one valid \(F \in \mathcal{F}\) for Algorithm 2 with parameter \(\beta\).

More precisely, let \(\sigma = \sigma_X(E) \oplus D\) then the set \(G := \{F \in \mathcal{F} : \Delta(\sigma, F) \geq \beta |\sigma_X(F)|\}\) satisfies:

\[
\sum_{F \in G} |\sigma_X(F)| \geq c_1 |\sigma_X(E)| - c_2 |D \cap \sigma_X(E)| > 0.
\]

**Lemma 11** (Robustness). We use the notations of Section III-A with \(\delta < 1/8\).

If \(E_R \subseteq V\) is a reduced error with \(|E_R| \leq \gamma_0 \sqrt{n}\) then:

\[
|\sigma_X(E_R)| \geq \frac{\beta_0 d_A}{2} |E_R|.
\]

Together, Proposition 10 and Lemma 11 imply the following:

**Corollary 12.** We use the notations of Section III-A with \(\delta < 1/16\) and \(\beta \in (0; \beta_1)\).

Suppose that \(E \oplus \bar{E}\) is reduced with \(|E \oplus \bar{E}| \leq \gamma_0 \sqrt{n}\) then:

\[
|E \oplus \bar{E}| \leq c_0 |D \cap \sigma_X(E \oplus \bar{E})|.
\]

**Proof:** Using the notations \(\sigma_i\) from the body of Algorithm 2, the value of the syndrome \(\sigma_{i+1}\) at the end of the algorithm is \(\sigma_{i+1} = \sigma_X(E \oplus \bar{E}) \oplus D\). Since the whole loop condition is not satisfied for \(\sigma_f\), the contraposition of Proposition 10 ensures that \(|\sigma_X(E \oplus \bar{E})| \leq c_0 |D \cap \sigma_X(E \oplus \bar{E})|\). By Lemma 11, \(|\sigma_X(E \oplus \bar{E})| \geq \frac{\beta_0 d_A}{2} |E \oplus \bar{E}|\) which concludes the proof.

The rest of Section III-C is devoted to prove Proposition 10 and Lemma 11.

We will study the sets \(F \in G (G\) is defined in Proposition 10) which would have been flipped during the small-set-flip algorithm with input \(E\) and \(D = \varnothing\), i.e., without syndrome error. For a given set \(E \subseteq V = A^2 \uplus B^2\) (where \(\uplus\) stands for disjoint union), we introduce a notation for a normalized Hamming weight:

\[
\|E\| := \frac{|E \cap A^2|}{d_B} + \frac{|E \cap B^2|}{d_A}.
\]

\(\|\cdot\|\) shares a couple of properties with the usual the cardinality \(|\cdot|\). In particular it is straightforward to check that for \(E, E_1, E_2 \subseteq V:\)

\[
\|E\| = 0 \iff E = \varnothing, \quad d_A \|E\| \leq |E| \leq d_B \|E\|, \\
|\sigma_X(E)| \leq d_A d_B \|E\|, \quad \|E_1 \cup E_2\| \leq \|E_1\| + \|E_2\|, \\
\|E_1 \cup E_2\| = \|E_1\| + \|E_2\|, \\
\|E_1 \cap E_2\| = \|E_1\| + \|E_2\| - 2 \|E_1 \cap E_2\|.
\]

We will say that a qubit error \(E \subseteq V\) is \(\|\cdot\|\)-reduced when \(\|E\|\) is minimal over \(E + C_z^2\). All along this section we will use the handy property of Lemma 13:

**Lemma 13.** Let \(E_1 \subseteq E_2 \subseteq V\) be two errors. If \(E_2\) is reduced (resp. \(\|\cdot\|\)-reduced) then \(E_1\) is reduced (resp. \(\|\cdot\|\)-reduced).

First of all, we need to study the case where the syndrome is noiseless \((D = \varnothing)\). In that case and when the initial graph \(G\) is sufficiently expanding, there exists at least one \(X\)-type stabilizer generator called a “critical generator” (this notion was introduced in [13]) whose support contains some \(F \in \mathcal{F}\) that decreases the syndrome weight when flipped.

**Lemma 14** (Lemma 8 of [13] revisited). Let \(E \subseteq V\) be a \(\|\cdot\|\)-reduced error such that \(0 < \|E\| \leq \gamma_0 \sqrt{n}/d_A\), then there exists \(F \in \mathcal{F}\) with \(F \subseteq E\) and:

(1) \(|\sigma_X(F)| \geq \frac{1}{2} d_A d_B \|F\|\),

(2) \(|\sigma_X(F)\| - \frac{\beta}{d_A} \|F\| \geq \sqrt{d_A} \|F\| - 4 \delta d_A d_B \|F\|\).

**Proof of Proposition 10 and Lemma 11:**

Both proofs begin in the same way: we set \(E_0\) to be the \(\|\cdot\|\)-reduced error equivalent to \(E\) (or equivalent to \(E_R\) in the case of Lemma 11), we apply Lemma 14 to \(E_0\) which provides some \(F_0 \subseteq E_0\) and we define \(E_1 \subseteq E_0 \oplus F_0 = E_0 \setminus F_0\). More generally, we set by induction \(E_{i+1} \subseteq E_i \oplus F_i = E_i \setminus F_i\) where \(F_i\) is obtained by applying Lemma 14 to \(E_i\).

This construction is licit (i.e., we can apply Lemma 14 to \(E_i\)) because \(E_i\) is \(\|\cdot\|\)-reduced as a subset of the \(\|\cdot\|\)-reduced error \(E_0\) (see Lemma 13), and \(\|E_i\| \leq \|E_0\| \leq \|E\|\). Let \(f'\) be the last step of this procedure then \(\|E_{f'}\| = 0\) and thus:

\[
E_0 = \biguplus_{i=0}^{f'-1} F_i.
\]

Since the sets \(F_i\) are those given in Lemma 14, we can use Lemma 14 items i and ii to lower bound \(|\sigma_X(E_0)|\):

\[
|\sigma_X(E_0)| = \sum_{i=0}^{f'-1} \Delta(\sigma_X(E_i), F_i) \geq \sum_{i=0}^{f'-1} |\sigma_X(F_i)| - \sum_{i=0}^{f'-1} 4 \delta d_A d_B \|F_i\| \geq \frac{\beta_0 d_A d_B}{2} \|E_0\|.
\]

Because \(E_0 = \biguplus F_i\) (eq. (5)), we have:

\[
|\sigma_X(E_0)| \geq \frac{\beta_0 d_A d_B}{2} \|E_0\|.
\]

The above arguments hold for Proposition 10 as well as for Lemma 11. Proving Lemma 11 is now direct:

\[
|\sigma_X(E_R)| = |\sigma_X(E_0)| \geq \frac{\beta_0 d_A d_B}{2} \|E_0\| \geq \frac{\beta_0 d_A}{2} |E_0| \geq \frac{\beta_0 d_A}{2} |E_R|.
\]

For Proposition 10, let us start by providing an overview of how we will proceed. Note that a union bound yields \(|\sigma_X(E)| \leq \sum_{i=0}^{f'-1} |\sigma_X(F_i)|\) in fact, we will prove in eq. (8) below that that this upper bound is nearly
tight: \(|\sigma_X(E)| \geq \beta_0 \sum_{i=0}^{f-1} |\sigma_X(F_i)|\) and \(\beta_0\) is arbitrarily close to 1 when \(\delta\) is small. Intuitively, this means that the intersection of the sets \(\sigma_X(F_i)\) is small and thus \(|\sigma_X(E) \cap \sigma_X(F_i)|\) is generally large. This is still true if the size of the syndrome error \(D\) is small, i.e., it holds that \(|\sigma \cap \sigma_X(F_i)|\) is generally large. Hence we will obtain Proposition 10 by computing the average of the quantity \(|\sigma \cap \sigma_X(F_i)|\) over the sets \(F_i\). We now provide the details:

\[
|\sigma_X(E)| \geq \sum_{i=0}^{f-1} |\sigma_X(F_i)| - \sum_{i=0}^{f-1} 4\delta d_A d_B \|F_i\| \quad \text{by eq. (6)}
\]

\[
= \sum_{i=0}^{f-1} |\sigma_X(F_i)| - 4\delta d_A d_B \|E_0\| \quad \text{by eq. (5)}
\]

\[
\geq \sum_{i=0}^{f-1} |\sigma_X(F_i)| - \frac{8\delta}{\beta_0} |\sigma_X(E)| \quad \text{by eq. (7)}.
\]

Hence we have:

\[
\left(1 - \frac{8\delta}{\beta_0}\right) |\sigma_X(E)| \geq \sum_{i=0}^{f-1} |\sigma_X(F_i)|. \quad \text{(8)}
\]

The relation between \(\Delta(\sigma, F)\) and \(|\sigma \cap \sigma_X(F)|\) is given by:

\(\Delta(\sigma, F) = |\sigma| - |\sigma \cap \sigma_X(F)| = 2|\sigma \cap \sigma_X(F)| - |\sigma_X(F)|\)

where we have used the equality \(|A_1 \oplus A_2| = |A_1| + |A_2| - 2|A_1 \cap A_2|\).

In particular, when \(F \notin G\):

\[
|\sigma \cap \sigma_X(F)| \leq \frac{1 + \beta}{2} |\sigma_X(F)|. \quad \text{(9)}
\]

On the one hand, eqs. (8) and (9) give an upper bound on the sum \(S := \sum_{i=0}^{f-1} |\sigma \cap \sigma_X(F_i)|\):

\[
S = \sum_{F_i \in G} |\sigma \cap \sigma_X(F_i)| + \sum_{F_i \notin G} |\sigma \cap \sigma_X(F_i)|
\]

\[
\leq \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1 + \beta}{2} \sum_{F_i \notin G} |\sigma_X(F_i)| \quad \text{by eq. (9)}
\]

\[
= \frac{1 - \beta}{2} \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1 + \beta}{2} \sum_{F_i \notin G} |\sigma_X(F_i)|
\]

\[
\leq \frac{1 - \beta}{2} \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1 + \beta}{2} \frac{1}{\beta_0} |\sigma_X(E)| \quad \text{by eq. (8)}.
\]

On the other hand, \(E_0 = \bigcup_{i=0}^{f} F_i\) (eq. (5)) implies that \(\sigma_X(E) = \sigma_X(E_0) = \bigoplus_{i=0}^{f-1} \sigma_X(F_i)\) and thus \(S\) is lower bounded by:

\[
S \geq |\sigma \cap \sigma_X(E)| = |(\sigma_X(E) \oplus D) \cap \sigma_X(E)|
\]

\[
= |\sigma_X(E) \oplus (D \cap \sigma_X(E))|
\]

\[
= |\sigma_X(E)| - |D \cap \sigma_X(E)|.
\]

Combining both inequalities we get Proposition 10.
by:

$$\text{MaxConn}_\alpha(E) = \max \{|X| : X \text{ is connected in } \mathcal{G} \}.$$

This notion of $\alpha$-subset is relevant because if we run the small-set-flip decoding algorithm of parameter $\beta > 0$ and set $U = E \cup F_0 \cup \ldots \cup F_{t-1}$ to be the execution support then $U \cup D$ is a $2\alpha_0$-subset of $E \cup D$.

Later we will need the following technical lemma in order to reduce to the case where the remaining error $E \oplus \hat{E}$ is reduced:

**Lemma 17.** Let $E, X_1, X_2 \subseteq V$ with $|X_2| \leq |X_1|$. If $X_1$ is an $\alpha$-subset of $E$ then $X_1 \cup X_2$ is an $\frac{\alpha}{2}$-subset of $E$.

**Proposition 18.** Using the notations of Section III-A, $U \cup D$ is a $2\alpha_0$-subset of $E \cup D$.

In order to prove Theorem 9, we would like to show that with probability at least $1 - e^{-\Omega(\sqrt{n})}$, there is a reduced error $E_{\alpha}$ equivalent to $E \oplus \hat{E}$ which is local stochastic.

Recall from Definition 7 that an error $E_{\alpha}$ is local stochastic with parameter $p$ if we can upper bound the probability $P[S \subseteq E_{\alpha}]$ by $p^{|S|}$ for all $S \subseteq V$. The reason why the probability $p$ provided by Theorem 9 depends on $p_{\text{synd}}$ is that we will use the hypothesis that $D$ is local stochastic: $P[T \subseteq D] \leq p_{\text{synd}}^{|T|}$ for all $T \subseteq C_X$. In order to establish Theorem 9, it would be sufficient to prove that for all $S \subseteq E_{\alpha}$, the size of $D \cap \Gamma_X(S)$ is lower bounded by a linear function of $|S|$. The particular case where $S = E_{\alpha} = E \oplus \hat{E}$ has already been proven in Corollary 12 but unfortunately it is not possible to prove this property for all $S \subseteq E_{\alpha}$.

Therefore, instead of lower bounding $|D \cap \Gamma_X(S)|$ linearly in $|S|$, we will lower bound $|D \cap \Gamma_X(W)|$ linearly in $|W|$ where $W$ contains $S$ and is such that we can apply the locality property of Lemma 15 to prove that Algorithm 2 cannot flip any set on input $(W, D \cap \Gamma_X(W))$. Applying Corollary 12 for the execution of Algorithm 2 on input $(W, D \cap \Gamma_X(W))$ will provide the desired lower bound. We will call such a $W$ a witness for $S$ (see Definition 20).

In the proof of Theorem 9, we will show that if we have witnesses for any $S \subseteq E_{\alpha}$ then $E_{\alpha}$ is local stochastic. This will require to be able to control the number of possible witnesses for a given $S$ and that is why we will need an additional condition $W \in \mathcal{M}(S)$ in the definition of a witness (see Definition 19 for $\mathcal{M}(S)$ and Definition 20 for witness).

**Definition 19 (4)].** Let $\mathcal{G} = (V \cup C_X, \mathcal{E})$ be the syndrome adjacency graph and $d := d_B(d_B + 2d_A - 1)$ an upper bound on the degrees in $\mathcal{G}$ (as defined above Lemma 15). For $S \subseteq V$, we define $\mathcal{M}(S) \subseteq \mathcal{P}(V)$ as the set of all subsets $W \subseteq V$ such that $S \subseteq W$ and such that any connected component $W'$ of $W$ in $\mathcal{G}$ satisfies $W' \cap S \neq \emptyset$.

In this definition, a set $W \in \mathcal{M}(S)$ is any superset of $S$ such that $S$ intersects every connected component of $W$. Lemma 2 of Ref. [4] provides an upper bound on the number of sets $W \in \mathcal{M}(S)$ of a given size:

$$|\{W \in \mathcal{M}(S) : |W| = t\}| \leq \frac{(ed)^t}{cd^{|S|}}.$$

**Definition 20.** Let $D \subseteq C_X$ be a syndrome error and $c$ be a constant. For $S \subseteq V$, we say that $W \subseteq V$ is a $c$-witness for $(S, D)$ if $W \in \mathcal{M}(S)$ and $|W| \leq c|D \cap \Gamma_X(W)|$.

The proofs are organized as follows: first we show that we can easily find a witness for any $S \subseteq E \oplus \hat{E}$ under the assumptions that $|E \oplus \hat{E}| \leq \gamma_0 \sqrt{n}$ and that $E \oplus \hat{E}$ is reduced (Lemma 21), second we use Lemma 21 in order to construct witnesses for $S \subseteq E_{\alpha}$ under the assumption that $|\text{MaxConn}_{\alpha}(E)| \leq \gamma_0 \sqrt{n}$ (Lemma 22), third we show using percolation arguments that $|\text{MaxConn}_{\alpha}(E)| \leq \gamma_0 \sqrt{n}$ holds with probability $1 - e^{-\Omega(\sqrt{n})}$ when $E$ and $D$ are local stochastic (Lemma 23) and finally we conclude the proof of Theorem 9 by showing that if there exist witnesses for all $S \subseteq E_{\alpha}$ then $E_{\alpha}$ is local stochastic.

**Lemma 21.** We use the notations of Section III-A with $\delta < 1/16$ and $\beta \in (0; \beta_1)$.

If the remaining error $E \oplus \hat{E}$ is reduced and $|E \oplus \hat{E}| \leq \gamma_0 \sqrt{n}$ then for all $S \subseteq E \oplus \hat{E}$, there is a $c_0$-witness $W$ for $(S, D)$ with the additional constraint $W \subseteq E \oplus \hat{E}$.

**Proof:** We set $E_R = E \oplus \hat{E}$ and define $W$ to be all the connected components of $E_R$ in $\mathcal{G}$ that contain at least one element of $S$. It is clear that $W \in \mathcal{M}(S)$ and that $W \subseteq E \oplus \hat{E}$ hold, and it remains to prove $|W| \leq c_0|D \cap \Gamma_X(W)|$.

By locality (Lemma 15 applied with $K = W$), no flip is done by Algorithm 2 on the input $E_R \wedge W, D \cap \Gamma_X(W)$. Moreover, the remaining error $(E \oplus \hat{E}) \cap W$ is reduced as a subset of the reduced error $E \oplus \hat{E}$ (Lemma 13) and satisfies $|E \oplus \hat{E}| \cap W| \leq \gamma_0 \sqrt{n}$. Hence Corollary 12 states that:

$$|W| = |(E \oplus \hat{E}) \cap W| \leq c_0|D \cap \Gamma_X(W) \cap \sigma_X((E \oplus \hat{E}) \cap W)| \leq c_0|D \cap \Gamma_X(W)|.$$

**Lemma 22.** We use the notations of Section III-A with $\delta < 1/16$ and $\beta \in (0; \beta_1)$.

If $|\text{MaxConn}_{\alpha}(E \cup D)| \leq \gamma_0 \sqrt{n}$ then there is a reduced error $E_{\alpha}$ equivalent to the remaining error $E \oplus \hat{E}$ such that for all $S \subseteq E_{\alpha}$ there is a $c_0$-witness for $(S, D)$.

The proof of this lemma is left to the full version [18].

The last ingredient before proving Theorem 9 is provided by Lemma 23 below: the condition $\text{MaxConn}_{\alpha}(E \cup D) \leq \gamma_0 \sqrt{n}$ needed in Lemma 22 is verified with high probability for local stochastic errors.

**Lemma 23** (alpha-percolation, [14]). Let $\mathcal{G} = (V \cup C_X, \mathcal{E})$ be the syndrome adjacency graph and $d := d_B(d_B + 2d_A - 1)$ a
1) an upper bound on the degrees in $G$ (as defined above Lemma 15). Then for any $\alpha \in (0, 1)$, there exists a threshold $p_h = p_h(\alpha, d) > 0$ such that for any $t \in \mathbb{N}^*$ and $p < p_h$:

$$
\Pr[\text{MaxConn}_\alpha(E) \geq t] \leq C|\mathcal{V}| \left( \frac{p}{p_h} \right)^{\alpha t},
$$

where $C = C(\alpha, p, p_h)$ is a constant and $E \subseteq \mathcal{V}$ is a random set chosen accordingly to a local stochastic noise of parameter $p$.

The proof of Theorem 9 follows by combining Lemma 22 and Lemma 23 and it is deferred to the full version of the paper [18].

ACKNOWLEDGMENT

We would like to thank Benjamin Audoux, Alain Couvreur, Aniruddh Krishna, Vivien Londe, Jean-Pierre Tillich and Gilles Zémor for many fruitful discussions on quantum codes as well as Daniel Gottesman for answering our questions about [4]. AG and AL acknowledge support from the ANR through the QuantERA project QCDA.

REFERENCES


