2018 IEEE 59th Annual Symposium on Foundations of Computer Science

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
omar.fawzi@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 $|\mathrm{C}|$ locations[1], $\mathcal{O}(\log\log(|\mathrm{C}|/\varepsilon))$ levels of encoding are needed, which translates into a $\mathrm{polylog}(|\mathrm{C}|/\varepsilon)$ 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 rate 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-

[1]A 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.

2575-8454/18/$31.00 ©2018 IEEE
DOI 10.1109/FOCS.2018.00076

tial preparation and final measurement, the encoded circuit alternates between steps applying a gate of the original circuit on the encoded state with Knill's technique, and steps of error correction for the quantum code consisting of a measurement of the syndrome, running a (sufficiently fast) classical decoding algorithm and applying the necessary correction. For this to work out properly, and to keep a constant overhead, the syndrome measurement should be efficient and this will be the case if the quantum code is a low-density parity-check (LDPC) code. Indeed, in that case, the syndrome measurement circuit will be of constant depth and won't require any additional overhead. The last property needed for the scheme to work is the existence of an efficient classical decoding algorithm for the quantum code working even in the presence of noise on the syndrome measurement.

The main result of Gottesman is as follows: provided that the right family of quantum codes exists, it is possible to perform fault-tolerant quantum computing with constant overhead. By right family, we mean a family of constant rate LDPC quantum codes with an efficient decoding algorithm robust against noisy syndrome measurements. More precisely, the decoding algorithm should correct stochastic errors of linear weight and fail with negligible probability. Ideally, we also want the algorithm to be sufficiently fast to avoid errors building up during the decoding. At the time, no such family of codes was known to exist. In fact, families of constant rate LDPC codes with unbounded minimum distance are quite difficult to construct, even forgetting about the decoding problem. Possible candidate families include surface codes [5], 4-dimensional topological codes [6] and hypergraph product codes [7]. While surface codes come with an efficient decoding algorithm based on minimum weight matching [8], they only display a logarithmic minimum distance if they have constant rate [9]. Topological 4-D codes come with a much larger minimum distance, but the available efficient decoding algorithms are only known to perform well for errors of logarithmic weight [10], [11]. In both cases, this is insufficient to provide a universal threshold for the error rate, independent of the size of the quantum circuit to be performed. Finally, the family of hypergraph product codes yields the best minimum distance to date for constant rate LDPC codes: the minimum distance scales like the square-root of the block length. In general, however, we don't know of any efficient decoding algorithm for hypergraph product codes.

The hypergraph product construction takes a classical code $[n, k, d_{\min}]$ as input an yields a quantum code $[[N, K, D_{\min}]]$ of length $N = \Theta(n^2)$, dimension $K = \Theta(k^2)$ and a minimum distance equal to that of the classical code. When applying this construction with a classical expander code [12], it yields a so-called *quantum expander code* [13]. In that case, one has $K = \Theta(N)$ and $D_{\min} = \Theta(\sqrt{N})$, and interestingly, one can take inspiration of the efficient bit-flip decoding algorithm for classical expander codes [12] to design an efficient decoding algorithm for the quantum expander codes. Such an algorithm, the *small-set-flip decoding algorithm*, was introduced in [13] where it was proved that it corrects arbitrary (adversarial) errors of weight $O(D_{\min})$ in linear time. More recently in [14], we studied the behavior of this algorithm against stochastic noise and proved that it corrects random errors of linear weight, except with negligible probability.

In the present work, we extend the analysis significantly and show that the algorithm still works in the presence of a noisy syndrome. This was the missing condition to satisfy all the criteria required by Gottesman's construction. In other words, quantum expander codes can be exploited to obtain quantum fault-tolerance with a constant overhead. In addition, we establish two remarkable features of the decoding algorithm: first, it is *single-shot* meaning that the syndrome measurement need not be repeated a polynomial number of times as in typical constructions [15]: one measurement suffices; second, the algorithm can be parallelized to run in logarithmic time instead of linear time. This second point is important since storage errors will always affect the qubits during the classical decoding step, meaning that it is crucial to reduce the necessary time as much as possible. We note, however, that for our main result below to hold, we need to assume that the error rate affecting the qubits during the decoding step is constant and doesn't depend on the size of the computation. Without this extra assumption in our model, it is implausible that true constant space overhead quantum fault-tolerance can be achieved.

We obtain the following general result by using our analysis of quantum expander codes in Gottesman's generic construction [4].

Theorem 1. *For any $\eta > 1$ and $\varepsilon > 0$, there exists $p_T(\eta) > 0$ such that the following holds for sufficiently large k. Let C be a quantum circuit acting on k qubits, and consisting of $f(k)$ locations for f an arbitrary polynomial. There exists a circuit $\tilde{\mathrm{C}}$ using ηk physical qubits, depth $\mathcal{O}(f(k))$ and number of locations $\mathcal{O}(kf(k))$ that outputs a distribution which has total variation distance at most ε from the output distribution of C, even if the components of $\tilde{\mathrm{C}}$ 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 σ 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 $(p_{\mathrm{phys}}, p_{\mathrm{synd}})$ with $p_{\mathrm{phys}} < p_0$ and $p_{\mathrm{synd}} < p_1$. 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_{ls} that has a local stochastic distribution with parameter $p_{\mathrm{synd}}^{\Omega(1)}$. In addition, the small-set-flip algorithm can be parallelized to run in $\mathcal{O}(\log n)$ depth.*

In the special case where the syndrome measurements are perfect, *i.e.*, $p_{\mathrm{synd}} = 0$, 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 $p_{\mathrm{synd}}^{\Omega(1)}$ (our choice of p_0 will be such that $p_{\mathrm{synd}}^{\Omega(1)} \ll p_0$). 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

$\{F_i\}$ to flip successively in order to correct the error. In other words, we obtain the following decomposition of the error, $E = \oplus_i F_i$ (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 $E \oplus \hat{E}$ 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 α-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 $E \oplus \hat{E}$ 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 $E \oplus \hat{E}$ 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 $\mathcal{C}$ 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: $\mathcal{C} = \{x \in \mathbb{F}_2^n : Hx = 0\}$. The minimum distance $d_{\min}$ of the code is the minimum Hamming weight of a nonzero codeword: $d_{\min} = \min\{|x| : x \in \mathcal{C}, x \neq 0\}$. 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 $G = (V \cup C, \mathcal{E})$ 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 $H_{cv} = 1$ if v and c are adjacent in G and $H_{cv} = 0$ otherwise. The binary linear code $\mathcal{C}_G$ associated with G is the code with parity-check matrix H. The graph G is the *factor graph* of the code $\mathcal{C}_G$, 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 $e \in \mathbb{F}_2^{n_V}$ 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:

$$\sigma(e) := He \in \mathbb{F}_2^{n_C}, \qquad \sigma(E) := \bigoplus_{v \in E} \Gamma(v) \subseteq C,$$

where $\Gamma(v) \subseteq C$ is the set of neighbors of v. In this paper, the operator $\oplus$ 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_V and d_C respectively. Let $|V| = n_V$ and $|C| = n_C$. We say that G is* (γ, δ)-left-expanding *for some constants $\gamma, \delta > 0$, if for any subset $S \subseteq V$ with $|S| \leq \gamma n_V$, 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* (γ, δ)-right-expanding *if for any subset $S \subseteq C$ with $|S| \leq \gamma n_C$, we have $|\Gamma(S)| \geq (1-\delta) d_C |S|$. Finally, the graph G is said* (γ, δ)-expanding *if it is both (γ, δ)-left expanding and (γ, δ)-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 $X = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right), Y = ZX, Z = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and I with an overall phase of ± 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 $\mathcal{C}_X$ and $\mathcal{C}_Z$ of length n satisfying $\mathcal{C}_Z^\perp \subseteq \mathcal{C}_X$ (or equivalently, $\mathcal{C}_X^\perp \subseteq \mathcal{C}_Z$), where the dual code $\mathcal{C}_X^\perp$ to $\mathcal{C}_X$ consists of the words which are orthogonal to all the words in $\mathcal{C}_X$. This condition also reads $H_X \cdot H_Z^T = 0$, if H_X and H_Z denote the respective parity-check matrices of $\mathcal{C}_X$ and $\mathcal{C}_Z$. The quantum code $CSS(\mathcal{C}_X, \mathcal{C}_Z)$ associated with $\mathcal{C}_X$ (used to correct X-type errors and corresponding to Z-type stabilizer generators) and $\mathcal{C}_Z$ (used to correct Z-type errors and corresponding to X-type stabilizer generators) has length n and is defined as the linear span of $\left\{ \sum_{z \in \mathcal{C}_Z^\perp} |x+z\rangle \, : \, x \in \mathcal{C}_X \right\}$, where $\{|x\rangle : x \in \mathbb{F}_2^n\}$ 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 $k = \dim(\mathcal{C}_X/\mathcal{C}_Z^\perp) = \dim(\mathcal{C}_Z/\mathcal{C}_X^\perp) = \dim \mathcal{C}_X + \dim \mathcal{C}_Z - n$. 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 $CSS(\mathcal{C}_X, \mathcal{C}_Z)$, one has $d_{\min} = \min(d_X, d_Z)$ where $d_X = \min\{|E| : E \in \mathcal{C}_X \setminus {\mathcal{C}_Z}^\perp\}$ and $d_Z = \min\{|E| : E \in \mathcal{C}_Z \setminus {\mathcal{C}_X}^\perp\}$. We say that $CSS(\mathcal{C}_X, \mathcal{C}_Z)$ 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 $\mathcal{C}_X$ (resp. of $\mathcal{C}_Z$). We will denote by Γ_X (resp. Γ_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 ρ to $p_{\mathbb{1}} \rho + p_X X \rho X + p_Y Y \rho Y + p_Z Z \rho Z$, for some $p_{\mathbb{1}}, p_X, p_Y, p_Z$. 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 $CSS(\mathcal{C}_X, \mathcal{C}_Z)$ 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 (σ_X, σ_Z) of syndromes and should return a pair of errors $(\hat{E}_X, \hat{E}_Z)$ such that $E_X + \hat{E}_X \in \mathcal{C}_Z^\perp$ and $E_Z + \hat{E}_Z \in \mathcal{C}_X^\perp$. 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 σ_X and returns $\hat{E}_X$ such that $\sigma_X(\hat{E}_X) = \sigma_X$, and a Z-decoding algorithm that takes as input σ_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 σ_X and σ_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 $G = (A \cup B, \mathcal{E})$ be a biregular (γ, δ)-expanding graph with δ 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 $d_A \geq \lceil \delta^{-1} \rceil$. Let $\mathcal{C}$ be the classical code associated with G, let $d_{\min}(\mathcal{C})$ be the minimal distance of $\mathcal{C}$ 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 $\mathcal{C}$ with itself admits the following parity check matrices:

$$H_X = \left(I_{n_A} \otimes H, H^T \otimes I_{n_B}\right),$$
$$H_Z = \left(H \otimes I_{n_A}, I_{n_B} \otimes H^T\right).$$

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* $[[n, k, d_{\min}]]$*, where* $n = {n_A}^2 + {n_B}^2, k \geq (n_A - n_B)^2$ *and* $d_{\min} = d_{\min}(\mathcal{C})$.

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 try to apply 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 $2^{d_A + d_B} = \mathcal{O}(1)$ possible patterns to examine for each generator.

Before describing the algorithm more precisely, let us introduce some additional notations. Let $\mathcal{X}$ be the set of subsets of V corresponding to X-type generators: $\mathcal{X} = \{\Gamma_Z(g) : g \in C_Z\} \subseteq \mathcal{P}(V)$, where $\mathcal{P}(V)$ is the power set of V. The indicator vectors of the elements of $\mathcal{X}$ span the dual code $\mathcal{C}_Z^{\perp}$. The condition for successful decoding of the X error E then asks that there exists a subset $X \subset \mathcal{X}$ 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 $\Gamma_Z(g)$ for some generator $g \in C_Z$ which decreases the syndrome weight $|\sigma|$. In other words, it tries to flip some element $F \in \mathcal{F}_0$ such that $\Delta(\sigma, F) > 0$ where:

$$\begin{aligned} \mathcal{F}_0 &:= \{F \subseteq \Gamma_Z(g) : g \in C_Z\}, \\ \Delta(\sigma, F) &:= |\sigma| - |\sigma \oplus \sigma_X(F)|. \end{aligned} \tag{1}$$

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-flip 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 $X \subseteq \mathcal{X}$, *i.e.* E and $\hat{E}$ are equivalent errors

$\hat{E}_0 = 0$; $\sigma_0 = \sigma$; $i = 0$
while $(\exists F \in \mathcal{F}_0 : \Delta(\sigma_i, F) > 0)$ **do**
  $F_i = \arg\max_{F \in \mathcal{F}_0} \dfrac{\Delta(\sigma_i, F)}{|F|}$
  $\hat{E}_{i+1} = \hat{E}_i \oplus F_i$
  $\sigma_{i+1} = \sigma_i \oplus \sigma_X(F_i)$ // $\sigma_{i+1} = \sigma_X(E \oplus \hat{E}_{i+1})$
  $i = i + 1$
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* $G = (A \cup B, \mathcal{E})$ *be a* (d_A, d_B)*-biregular* (γ, δ)*-expanding graph*

[3]The existence of an efficient algorithm that corrects arbitrary errors of size $O(\sqrt{n})$ is guaranteed as soon as $\delta < 1/6$ [13]. The same algorithm corrects random errors of linear weight except with negligible probability as soon as $\delta < 1/8$ [14] and in the present paper, we will require the more stringent condition $\delta < 1/16$.

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 $w_0 = \frac{\gamma n_B}{3(1+d_B)}$.

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, Grospellier, Leverrier [14])**.** *Let $G = (A \cup B, \mathcal{E})$ be a (d_A, d_B)-biregular (γ, δ)-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 $1 - Cn\left(\frac{p}{p_0}\right)^{C'\sqrt{n}}$.*

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

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

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 σ_X, then proceeds in a similar way to try to recover E_Z from σ_Z, without exploiting any information about E_X or σ_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

$$\mathbb{P}[S \subseteq E, T \subseteq D] \leq p^{|S|} q^{|T|}. \tag{3}$$

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 $\sigma = \sigma_X(E)$, the 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 $\mathcal{F}$ instead of $\mathcal{F}_0$ as set of possible flips (see Remark 8 for a discussion about these changes):

$$\mathcal{F} := \left\{ F \subseteq \Gamma_Z(g) : g \in C_Z \text{ and } |\sigma_X(F)| \geq \frac{d_A}{2}|F| \right\}. \tag{4}$$

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)

$\hat{E}_0 = 0$; $\sigma_0 = \sigma$; $i = 0$
while $\exists F_i \in \mathcal{F} : \Delta(\sigma_i, F_i) \geq \beta|\sigma_X(F_i)|$ **do**
  $\hat{E}_{i+1} = \hat{E}_i \oplus F_i$
  $\sigma_{i+1} = \sigma_i \oplus \sigma_X(F_i)$ // $\sigma_{i+1} = \sigma_X(E \oplus \hat{E}_{i+1}) \oplus D$
  $i = i + 1$
end while
return $\hat{E}_i$

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 remaining error, we will 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 $|\sigma_X(F)| \geq \frac{d_A}{2}|F|$ 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 $D = \varnothing$ 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 $D \neq \varnothing$, requiring $\Delta(\sigma_i, F_i) \geq \beta|\sigma_X(F_i)|$ with β close to 1 makes Algorithm 2 more robust against syndrome errors.

The behavior of Algorithm 2 in the particular case where $D = \varnothing$ could be studied following the proof of [14]: given $\mathcal{Q}_G$ a quantum expander code constructed using G some bipartite (γ, δ)-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 β_0 (β_0 as defined in Section III-A). In the noisy case $D \neq \varnothing$, 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 $\beta < \beta_1$ (β_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 $G = (A \cup B, \mathcal{E})$ a (d_A, d_B)-biregular (γ, δ)-expander graph with $\gamma > 0$, we denote by $\mathcal{Q}$ 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 $\mathcal{Q}$. We will also use Γ_X and Γ_Z the neighborhoods in the graphs G_X ans G_Z as defined in Section II-C.
We run the small-set-flip decoding algorithm (Algorithm 2) of parameter β 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:

$$r := d_A/d_B, \qquad \gamma_0 = \frac{r^2}{\sqrt{1+r^2}}\gamma,$$

$$\beta_0 = \beta_0(\delta) := 1 - 8\delta, \qquad \beta_1 = \beta_1(\delta) := 1 - 16\delta,$$

$$c_0 = c_0(\delta, \beta) := \frac{4}{d_A(\beta_1 - \beta)}, \quad c_2 = c_2(\delta, \beta) := \frac{2\beta_0}{\beta_1 - \beta},$$

$$c_1 = c_1(\delta, \beta) := \frac{\beta_1 - \beta}{\beta_0(1 - \beta)}, \quad \alpha_0 = \alpha_0(\beta) := \frac{r\beta}{4 + 2r\beta}.$$

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 succ *that has probability $1 - e^{-\Omega(\sqrt{n})}$ and a random variable E_{ls} that is equivalent to $E \oplus \hat{E}$ such that conditioned on* succ*, E_{ls} has a local stochastic distribution with parameter Kp_{synd}^{1/c_0} where K is a constant independent of p_{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 $|D \cap \sigma_X(E \oplus \hat{E})|$. 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 β.
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:

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

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 $E \oplus \hat{E}$ is reduced with $|E \oplus \hat{E}| \leq \gamma_0\sqrt{n}$ then

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

Proof: Using the notations σ_i from the body of Algorithm 2, the value of the syndrome σ_f at the end of the algorithm is $\sigma_f = \sigma_X(E \oplus \hat{E}) \oplus D$. Since the while loop condition is not satisfied for σ_f, the contraposition of Proposition 10 ensures that $|\sigma_X(E \oplus \hat{E})| \leq c_2|D \cap \sigma_X(E \oplus \hat{E})|$. By Lemma 11, $|\sigma_X(E \oplus \hat{E})| \geq \frac{\beta_0 d_A}{2}|E \oplus \hat{E}|$ 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 $E, E_1, E_2 \subseteq V$:

$$\|E\| = 0 \Leftrightarrow E = \varnothing, \qquad d_A\|E\| \leq |E| \leq d_B\|E\|,$$
$$|\sigma_X(E)| \leq d_A d_B\|E\|, \qquad \|E_1 \cup E_2\| \leq \|E_1\| + \|E_2\|,$$
$$\|E_1 \uplus E_2\| = \|E_1\| + \|E_2\|,$$
$$\|E_1 \oplus E_2\| = \|E_1\| + \|E_2\| - 2\|E_1 \cap E_2\|.$$

We will say that a qubit error $E \subseteq V$ is $\|\ \|$-reduced when $\|E\|$ is minimal over $E + {\mathcal{C}_Z}^\perp$. 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. $\|\ \|$-reduced) then E_1 is reduced (resp. $\|\ \|$-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 $\|\ \|$-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:*
(i) $|\sigma_X(F)| \geq \frac{1}{2}d_A d_B\|F\|$,
(ii) $\Delta(\sigma_X(E), F) \geq |\sigma_X(F)| - 4\delta d_A d_B\|F\|$.

Proof of Proposition 10 and Lemma 11.:
Both proofs begin in the same way: we set E_0 to be the $\|\ \|$-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 $E_1 := E_0 \oplus F_0 = E_0 \setminus F_0$. More generally, we set by induction $E_{i+1} = E_i \oplus F_i = E_i \setminus F_i$ 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 $\|\ \|$-reduced as a subset of the $\|\ \|$-reduced error E_0 (see Lemma 13), and $\|E_i\| \leq \|E_0\| \leq \|E\| \leq |E|/d_A \leq \gamma_0\sqrt{n}/d_A$. Let f' be the last step of this procedure then $\|E_{f'}\| = 0$ and thus:

$$E_0 = \biguplus_{i=0}^{f'-1} F_i. \tag{5}$$

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

$$\begin{aligned} |\sigma_X(E_0)| &= \sum_{i=0}^{f'-1} \Delta(\sigma_X(E_i), F_i) \\ &\geq \sum_{i=0}^{f'-1} |\sigma_X(F_i)| - \sum_{i=0}^{f'-1} 4\delta d_A d_B\|F_i\| \\ &\geq \frac{\beta_0 d_A d_B}{2} \sum_{i=0}^{f'-1} \|F_i\|. \end{aligned} \tag{6}$$

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

$$|\sigma_X(E_0)| \geq \frac{\beta_0 d_A d_B}{2}\|E_0\|. \tag{7}$$

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

$$\begin{aligned} |\sigma_X(E_\mathrm{R})| &= |\sigma_X(E_0)| \\ &\geq \frac{\beta_0 d_A d_B}{2}\|E_0\| \geq \frac{\beta_0 d_A}{2}|E_0| \geq \frac{\beta_0 d_A}{2}|E_\mathrm{R}|. \end{aligned}$$

For Proposition 10, let us start by providing an overview of how we will proceed. Note that a union bound yields $|\sigma_X(E)| = |\sigma_X(E_0)| \leq \sum_{i=0}^{f'-1} |\sigma_X(F_i)|$. 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 β_0 is arbitrarily close to 1 when δ 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:

$$\begin{aligned}
|\sigma_X(E)| &\geq \sum_{i=0}^{f'-1} |\sigma_X(F_i)| - \sum_{i=0}^{f'-1} 4\delta d_A d_B \|F_i\| \quad \text{by eq. (6)} \\
&= \sum_{i=0}^{f'-1} |\sigma_X(F_i)| - 4\delta d_A d_B \|E_0\| \quad \text{by eq. (5)} \\
&\geq \sum_{i=0}^{f'-1} |\sigma_X(F_i)| - \frac{8\delta}{\beta_0} |\sigma_X(E)| \quad \text{by eq. (7).}
\end{aligned}$$

Hence we have:

$$\left(1 + \frac{8\delta}{\beta_0}\right) |\sigma_X(E)| \geq \sum_{i=0}^{f'-1} |\sigma_X(F_i)|. \tag{8}$$

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

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

where we have used the equality $|A_1 \oplus A_2| = |A_1| + |A_2| - 2|A_1 \cap A_2|$.
In particular, when $F \notin G$:

$$|\sigma \cap \sigma_X(F)| \leq \frac{1+\beta}{2} |\sigma_X(F)|. \tag{9}$$

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

$$\begin{aligned}
S &= \sum_{F_i \in G} |\sigma \cap \sigma_X(F_i)| + \sum_{F_i \notin G} |\sigma \cap \sigma_X(F_i)| \\
&\leq \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1+\beta}{2} \sum_{F_i \notin G} |\sigma_X(F_i)| \quad \text{by eq. (9)} \\
&= \frac{1-\beta}{2} \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1+\beta}{2} \sum_{i=0}^{f'-1} |\sigma_X(F_i)| \\
&\leq \frac{1-\beta}{2} \sum_{F_i \in G} |\sigma_X(F_i)| + \frac{1+\beta}{2\beta_0} |\sigma_X(E)| \quad \text{by eq. (8).}
\end{aligned}$$

On the other hand, $E_0 = \biguplus_i F_i$ (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:

$$\begin{aligned}
S &\geq |\sigma \cap \sigma_X(E)| \\
&= |(\sigma_X(E) \oplus D) \cap \sigma_X(E)| \\
&= |\sigma_X(E) \oplus (D \cap \sigma_X(E))| \\
&= |\sigma_X(E)| - |D \cap \sigma_X(E)|.
\end{aligned}$$

Combining both inequalities we get Proposition 10. ∎

D. Random errors of linear size

The upper bound given by Proposition 10 can be applied for qubit errors of size up to $t = \mathcal{O}(\sqrt{n})$. In the case of a local stochastic noise, the errors have a typical size of $\Theta(n)$. The relationship between the two frameworks is given by percolation arguments close to the arguments used in [14].

Percolation arguments will allow us to decompose a random error E as a disjoint union of small error sets, each of which has size upper bounded by t. The study of these small errors has been done in Section III-C and a consequence of Corollary 12 is that when we use Algorithm 2 to correct it, the remaining error is local stochastic. Moreover, Algorithm 2 is intuitively local in the sense that two qubit errors far away in the factor graph of the code will not interact during the decoding procedure. Under the assumption that a local stochastic error is produced for each small error set, the locality property will allow us to conclude that when we correct the initial error E with Algorithm 2, the remaining error has a local stochastic distribution.
In order to formalize the notion of locality, we define $\mathcal{G}$ called the syndrome adjacency graph of the code in the following way: $\mathcal{G}$ is equal to G_X (as defined in Section II-C) with additional edges between the qubits which share an X-type or a Z-type generator. In other words, the set of vertices of $\mathcal{G}$ is indexed by $\mathcal{V} := V \cup C_X$ the set of qubits and the set of Z-type generators, a Z-type generator is incident to the qubits in its support and two qubits are linked when they are both in the support of the same generator. Note that the degree of $\mathcal{G}$ is upper bounded by $d := d_B(d_B + 2d_A - 1)$. Using the execution support $U = E \cup F_0 \cup \ldots \cup F_{f-1}$, it is easy to decompose the error into small error sets: each connected component K of U provides one error set $K \cap E$.

Lemma 15 (Locality of Algorithm 2, [14])**.** *We use the notations of Section III-A.*
For any set $K \subseteq V$ with $\Gamma(K) \cap \Gamma(U \setminus K) = \varnothing$ in $\mathcal{G}$, there is a valid execution of Algorithm 2 on the input $(E \cap K, D \cap \Gamma_X(K))$ whose output is $\hat{E} \cap K$ and whose support is $U \cap K$.

What is the size of the remaining error $(E \oplus \hat{E}) \cap K$? We will show that this size is small enough to apply Corollary 12. The key point is to note that among the vertices of $X := K \cup (D \cap \Gamma_X(K)) \subseteq \mathcal{V}$, there is at least a fraction $2\alpha_0$ of these vertices which belong to $E \cup D$ (α_0 is defined in Section III-A). We will say that X is a $2\alpha_0$-subset of $E \cup D$ (see Definition 16) and percolation arguments (see Lemma 23) will show that with high probability, any connected α-subset of a random error $E \cup D$ must be small enough to apply Corollary 12.

Definition 16 ([14])**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, let $\alpha \in (0; 1]$ and let $E, X \subseteq \mathcal{V}$. X is said to be an α-*subset of E *if $|X \cap E| \geq \alpha |X|$. We also define the integer MaxConn$_\alpha(E)$*

by:

$$\textit{MaxConn}_\alpha(E) = \max\{|X| : X \textit{ is connected in } \mathcal{G} \\ X \textit{is an } \alpha\textit{-subset of } E\}.$$

This notion of α-subset is relevant because if we run the small-set-flip decoding algorithm of parameter $\beta > 0$ and set $U = E \cup F_0 \cup \ldots \cup F_{f-1}$ 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 $E, X_1, X_2 \subseteq \mathcal{V}$ with $|X_2| \leq |X_1|$. If X_1 is an α-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_{ls} equivalent to $E \oplus \hat{E}$ which is local stochastic.

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

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 $S \subseteq E_{\text{ls}}$ then E_{ls} 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 $W' \cap S \neq \varnothing$.*

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:

$$|\{W \in \mathcal{M}(S) : |W| = t\}| \leq \frac{(ed)^t}{ed^{|S|}}. \tag{10}$$

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 follow: 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 $S \subseteq E_{\text{ls}}$ under the assumption that $|\text{MaxConn}_{\alpha_0}(E)| \leq \gamma_0\sqrt{n}$ (Lemma 22), third we show using percolation arguments that $|\text{MaxConn}_{\alpha_0}(E)| \leq \gamma_0\sqrt{n}$ 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 $S \subseteq E_{\text{ls}}$ then E_{ls} 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 $E_{\text{R}} = E \oplus \hat{E}$ 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 $E_{\text{R}} \cap W, D \cap \Gamma_X(W)$. 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:

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

■

Lemma 22. *We use the notations of Section III-A with $\delta < 1/16$ and $\beta \in (0; \beta_1)$.*
If $|\textit{MaxConn}_{\alpha_0}(E \cup D)| \leq \gamma_0\sqrt{n}$ then there is a reduced error E_{ls} equivalent to the remaining error $E \oplus \hat{E}$ such that for all $S \subseteq E_{ls}$ 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 $\text{MaxConn}_{\alpha_0}(E \cup D) \leq \gamma_0\sqrt{n}$ needed in Lemma 22 is verified with high probability for local stochastic errors.

Lemma 23 (α-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) *an upper bound on the degrees in $\mathcal{G}$ (as defined above Lemma 15). Then for any $\alpha \in (0,1]$, there exists a threshold $p_{th} = p_{th}(\alpha, d) > 0$ such that for any $t \in \mathbb{N}^*$ and $p < p_{th}$:*

$$\mathbb{P}[\mathit{MaxConn}_\alpha(E) \geq t] \leq C|\mathcal{V}| \left(\frac{p}{p_{th}}\right)^{\alpha t},$$

where $C = C(\alpha, p, p_{th})$ 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, Anirudh 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

[1] D. Aharonov and M. Ben-Or, "Fault-tolerant quantum computation with constant error rate," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1207–1282, 2008.

[2] R. Reischuk and B. Schmeltz, "Reliable computation with noisy circuits and decision trees-a general n log n lower bound," in *Foundations of Computer Science, 1991. Proceedings., 32nd Annual Symposium on*. IEEE, 1991, pp. 602–611.

[3] P. Gács and A. Gál, "Lower bounds for the complexity of reliable boolean circuits with noisy gates," *IEEE Transactions on Information Theory*, vol. 40, no. 2, pp. 579–583, 1994.

[4] D. Gottesman, "Fault-tolerant quantum computation with constant overhead," *Quantum Information & Computation*, vol. 14, no. 15-16, pp. 1338–1372, 2014.

[5] M. H. Freedman, D. A. Meyer, and F. Luo, "Z2-systolic freedom and quantum codes," *Mathematics of quantum computation, Chapman & Hall/CRC*, pp. 287–320, 2002.

[6] L. Guth and A. Lubotzky, "Quantum error correcting codes and 4-dimensional arithmetic hyperbolic manifolds," *Journal of Mathematical Physics*, vol. 55, no. 8, p. 082202, 2014.

[7] J.-P. Tillich and G. Zémor, "Quantum LDPC codes with positive rate and minimum distance proportional to the square root of the blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1193–1202, 2014.

[8] J. Edmonds, "Maximum matching and a polyhedron with 0, 1-vertices," *Journal of Research of the National Bureau of Standards B*, vol. 69, no. 125-130, pp. 55–56, 1965.

[9] N. Delfosse, "Tradeoffs for reliable quantum information storage in surface codes and color codes," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 917–921.

[10] M. B. Hastings, "Decoding in hyperbolic spaces: quantum LDPC codes with linear rate and efficient error correction," *Quantum Information & Computation*, vol. 14, no. 13-14, pp. 1187–1202, 2014.

[11] V. Londe and A. Leverrier, "Golden codes: quantum LDPC codes built from regular tessellations of hyperbolic 4-manifolds," *arXiv preprint arXiv:1712.08578*, 2017.

[12] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1710–1722, 1996.

[13] A. Leverrier, J.-P. Tillich, and G. Zémor, "Quantum expander codes," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 810–824.

[14] O. Fawzi, A. Grospellier, and A. Leverrier, "Efficient decoding of random errors for quantum expander codes," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 521–534.

[15] H. Bombín, "Single-shot fault-tolerant quantum error correction," *Physical Review X*, vol. 5, no. 3, p. 031043, 2015.

[16] A. A. Kovalev and L. P. Pryadko, "Fault tolerance of quantum low-density parity check codes with sublinear distance scaling," *Physical Review A*, vol. 87, no. 2, p. 020304, 2013.

[17] E. T. Campbell, "A theory of single-shot error correction for adversarial noise," *arXiv preprint arXiv:1805.09271*, 2018.

[18] O. Fawzi, A. Grospellier, and A. Leverrier, "Constant overhead quantum fault-tolerance with quantum expander codes," *arXiv preprint arXiv:1808.03821*, 2018.

[19] R. Gallager, "Low-density parity-check codes," *IRE Transactions on information theory*, vol. 8, no. 1, pp. 21–28, 1962.

[20] D. Gottesman, "Stabilizer codes and quantum error correction," Ph.D. dissertation, California Institute of Technology, 1997.

[21] A. R. Calderbank and P. W. Shor, "Good quantum error-correcting codes exist," *Physical Review A*, vol. 54, no. 2, p. 1098, 1996.

[22] A. M. Steane, "Error correcting codes in quantum theory," *Physical Review Letters*, vol. 77, no. 5, p. 793, 1996.

[23] D. Gottesman, "An introduction to quantum error correction and fault-tolerant quantum computation," *arXiv preprint arXiv:0904.2557*, 2009.