Efficient estimation of Pauli channels

Steven T. Flammia[1, 2, 3] and Joel J. Wallman[3, 4]

[1] *Centre for Engineered Quantum Systems, School of Physics,*
University of Sydney, Sydney, NSW 2006 Australia
[2] *Yale Quantum Institute, Yale University, New Haven, CT 06520, USA*
[3] *Quantum Benchmark Inc., 100 Ahrens Street West,*
Suite 203, Kitchener, ON N2H 4C3, Canada
[4] *Institute for Quantum Computing and Department of Applied Mathematics,*
University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

(Dated: July 31, 2019)

Pauli channels are ubiquitous in quantum information, both as a dominant noise source in many computing architectures and as a practical model for analyzing error correction and fault tolerance. Here we prove several results on efficiently learning Pauli channels, and more generally the Pauli projection of a quantum channel. We first derive a procedure for learning a Pauli channel on n qubits with high probability to a *relative* precision ϵ using $O\big(\epsilon^{-2}n2^n\big)$ measurements, which is efficient in the Hilbert space dimension. The estimate is robust to state preparation and measurement errors which, together with the relative precision, makes it especially appropriate for applications involving characterization of high-accuracy quantum gates. Next we show that the error rates for an arbitrary set of s Pauli errors can be estimated to a relative precision ϵ using $O\big(\epsilon^{-4}\log s\log s/\epsilon\big)$ measurements. Finally, we show that when the Pauli channel is given by a Markov field with at most k-local correlations, we can learn an entire n-qubit Pauli channel to relative precision ϵ with only $O_k\big(\epsilon^{-2}n^2\log n\big)$ measurements, which is efficient in the number of qubits. These results enable a host of applications beyond just characterizing noise in a large-scale quantum system: they pave the way to tailoring quantum codes, optimizing decoders, and customizing fault tolerance procedures to suit a particular device.

I. INTRODUCTION

Pauli channels are among the most basic noise channels in quantum information science. Nearly all theoretical studies of quantum error correction and fault tolerance, including most threshold and overhead estimates, rely on modeling noise as a Pauli channel [1].

Originally, the theoretical focus on Pauli channels was primarily motivated by the ease with which they can be theoretically analyzed and simulated on a classical computer. However, Pauli channels are now justified by the technique of randomized compiling [2], which maps general quantum noise to its Pauli projection, a Pauli channel having the same average fidelity to an ideal channel. This addresses the concern that coherent noise, which is in general as hard to model as full quantum computation, may create distorted comparisons with threshold error rates that were computed using Pauli noise [3, 4]. Moreover, quantum error correction of coherent noise will, under reasonable assumptions, lead to less coherent noise at the logical level, meaning that noise at that level is better approximated by a Pauli channel [5, 6]. Finally, recent experiments [7] have shown that enforcing the Pauli projection by randomized compiling works extremely well in practice, further solidifying the importance of Pauli channels.

Despite the central role played by Pauli channels, to date there have been few systematic studies of how to estimate them efficiently, meaning with a complexity that improves over what follows from a naive application of full channel tomography. We will review the relevant literature below. Filling this gap becomes even more pressing in light of work showing just how much the threshold changes under biased or correlated noise models [8–11], in some cases by more than a factor of 4 for a code capacity threshold [12, 13]. Such factors can be substantial because the logical error rate depends exponentially on the distance below the threshold. Accurate estimation of the Pauli noise in an architecture would allow for many applications, including tailoring codes and decoders to match the noise [14], customizing fault-tolerance schemes [15], and accurate estimation of thresholds and overheads [16].

A. Summary of Results

In this paper, we give a comprehensive treatment of the sample complexity of Pauli channel estimation. We will present three main results corresponding to efficient estimation procedures with error guarantees for three separate and complementary regimes: estimation of a complete Pauli channel, estimation of error rates for an arbitrary set of Pauli errors, and estimation of a Pauli channel that factorizes over a bounded-degree factor graph. The latter is equivalent to considering the Pauli error probability as a tensor network state comprised of tensors with a bounded number of indices, but no restriction on the topology of the connections. This section contains only informal statements of our results; the precise versions of these statements along with rigorous proofs are given in the subsequent sections.

The basic procedure that we develop and analyze is a variant of randomized benchmarking [17–19] and its recently introduced cousins, character benchmarking [20] and cycle benchmarking [21]. The procedure, defined in Section III, inherits the robustness to errors in state preparation and measurement (SPAM) enjoyed by all benchmarking variants. It uses preparations and measurements in a stabilizer basis (or equivalently, in the computational basis together with a single Clifford group element), and repeated rounds of random Pauli gates to average over the noise. Our results apply when the noise on the random Pauli gates, the state preparations, and the measurement all obey certain mild regularity assumptions such as being gate-independent, time-stationary, Markovian, and not too far from ideal. These assumptions can be relaxed still further, particularly the gate-independence assumption [22–25]. The precise conditions on the noise for which our proofs hold are given in Definitions 2 and 3. In this section, we informally refer to these restrictions as "nice" noise.

Our first main result is that our procedure for estimating a complete Pauli channel on n qubits requires only $O\big(\epsilon^{-2}n2^n\big)$ measurements to estimate the channel to constant *relative* precision ϵ and with a constant success probability. This result forms a core subroutine in our subsequent two results. The output is the vector $\boldsymbol{p}$ of all 4^n Pauli error probabilities, which we achieve by using the full power of n-bit measurements. Here the relative precision and robustness to SPAM are crucial for applications, since average error rates in quantum gates are now routinely below 1%, and in some cases as low as 10^{-6} [26]. In this regime, a meaningful additive error approximation would require at least 10^{12} samples, making it far outside the realm of practicality even if it is still technically "efficient".

Result 1 (Informal summary of Propositions 8 and 9)**.** *For a nice Pauli channel on n qubits, the Pauli error rates $\boldsymbol{p}$ can be estimated using $O\big(\epsilon^{-2}n2^n\big)$ measurements by $\hat{\boldsymbol{p}}$ such that*

$$\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_2 \leq O(\epsilon)(1-p_0),$$

holds with high probability, where p_0 is the probability of no error.

We can apply a randomized sampling routine to estimate a subnormalized probability distribution over an arbitrary set E of size $|\mathsf{E}| = s$. We note that the ϵ^{-4} scaling is an artifact of the proof technique and expect it can be improved to ϵ^{-2} by a more careful analysis of the bias in the output of the subroutine **Ratio** described in section IV.

Result 2 (Informal statement of Theorem 11)**.** *For a nice Pauli channel on n qubits, a subnormalized distribution over any set E of s Pauli errors can be estimated using $O(\epsilon^{-4}\log(s)\log(s/\epsilon^2))$ measurements by $\hat{\boldsymbol{p}}$ such that*

$$\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_\infty \leq O(\epsilon)(1-p_0)$$

holds with constant probability, where p_0 is the probability of no error.

The above sampling protocol can be applied directly to efficiently estimate the probability of all low-weight Pauli errors, which will account for the majority of the distribution under realistic physical assumptions in near-term high-performance quantum systems. We also provide a tree-based search heuristic that can be used to identify sets of errors of interest. Two natural candidates for interesting sets of errors are the best s-sparse approximation without assuming local errors, and high-weight Pauli errors that occur with probabilities that differ substantially from model predictions. Although we do not give a formal analysis of the probability with which the search heuristic identifies a correct set, when it finds a sparse set with large measure it is certifiably correct as a description of the Pauli channel. This follows because theorem 11 can be applied to the final output without accounting for failure probabilities from intermediate steps.

Our final main result is the efficient reconstruction of any nice n-qubit Pauli channel in polynomial time in n whenever the channel has bounded-degree correlations and the local marginals are positive, conditions which we quantify in section VII. We assume throughout that the correlations are modeled by a factor graph with fixed known topology.

Result 3 (Informal statement of Proposition 18.)**.** *Let $\boldsymbol{p}$ be a nice n-qubit Pauli channel with a k-degree factor graph and positive marginals. Then an estimate $\hat{\boldsymbol{p}}$ as a tensor network can be obtained using $O_k\left(\epsilon^{-2}n^2\log n\right)$ measurements such that*

$$\|\hat{\boldsymbol{p}}-\boldsymbol{p}\|_1 \le O(\epsilon)\|\mathbf{1}_I-\boldsymbol{p}\|_\infty$$

holds with high probability. Moreover, an estimate proportional to $\hat{\boldsymbol{p}}$ can be found in time poly(n).

B. Proof techniques

As mentioned above, we make use of a variant of randomized benchmarking (RB). RB was originally applied to the full unitary group and the Clifford group [17, 18], but subsequent work has developed it for many different groups [27–34]; see Refs. [20, 35] for a general treatment. Our work focuses on RB over the Pauli group, and introduces several important advances over prior art.

As shown by Harper et al. [36], RB over the Clifford group provably gives relative precision estimates of a single parameter associated to a quantum channel, the average error rate. Our proof extends this to the case of RB over the Pauli group. The original proof of Ref. [36] isolates a single exponential decay from an RB signal using the method from Ref. [37]. However, for the case of interest here, the Pauli group, there are 4^n parameters that must be learned instead of just one. While a general solution exists to the problem of isolating exponential decays [20], prior work estimates the individual decay rates serially using one-bit measurements. Our work greatly improves this by showing how to take the natural n-bit measurements arising from measuring the n individual qubits and using these measurements to estimate up to 2^n parameters in parallel. When sampled in this way, a Hadamard transformation can be done to isolate the exponentials on all data collected from the n-bit measurements simultaneously, avoiding the need for serial exponential fits. Our proof for the case of estimating probabilities for arbitrary sets of Pauli errors uses the essential ideas above with random sampling techniques.

Finally, we assume that the Pauli channel error rates are given by a Markov random field over a known factor graph. We can again apply the core subroutine for estimating Pauli channels to the marginal channel on each of the factors and then round these estimated marginals into a coherent global probability distribution using the Hammersley-Clifford theorem [38].

C. Comparison with Prior Work

These results comprise the first proofs in the literature of recovery guarantees for quantum channels to relative precision while avoiding bias from SPAM errors. As noted above, this is absolutely essential for applications. Furthermore, Results 2 & 3 are the first recovery guarantees known that apply to broad classes of non-trivial quantum channels on n qubits with a scaling that is efficient in n (when $s = \text{poly}(n)$ or $k = O(1)$, respectively). Lastly, although numerical practicality is not a focus of the present paper (it will be explored elsewhere [39]) we wish to stress that these results are *practical* and immediately applicable to characterizing error rates in near-term devices, such as the 50- and 72-qubit processors announced by IBM and Google respectively. To emphasize this last point, the results of our numerical simulations of these procedures on up to 100 qubits show excellent performance, and we have successfully implemented a variant of the method on a publicly accessible 14-qubit quantum device [39]. To put this in perspective, the previous record for quantum channel tomography is of a 3-qubit quantum gate [40], which was only made possible by employing sophisticated methods from compressed sensing [41–43]. It is clear from these examples that our results represent a qualitative shift in the characterization of quantum devices.

Let us compare the scaling of our procedures to the best previously known results from the literature. A naive, non-adaptive application of channel tomography with (single-copy) two-outcome measurements [44] would require $\tilde{O}\big(\frac{d^4}{\epsilon^2}\big)$ state preparations and measurements to achieve an *additive* precision ϵ in, say, the average fidelity of the channel, where $d = 2^n$ for n qubits. Applying the best known procedures for state tomography [45–47] to the Choi state of the channel using a collective measurement requires $\tilde{\Theta}\big(\frac{d^4}{\epsilon^2}\big)$ copies, so this is roughly tight. For sparse channels in a known superoperator basis (such as the Pauli basis), Ref. [43] argues from the theory of compressed sensing that $O(s \log d)$ random product basis measurements should suffice to reconstruct an s-sparse channel in the Pauli basis, but it is not clear that the technical "incoherence" or restricted isometry properties that are used for efficiently reconstructing sparse vectors are satisfied in this case [42]. The recently introduced procedure of shadow tomography [48] could achieve an estimate of the Pauli error rates with a small ($\operatorname{poly}\log d$) number of copies of the Choi matrix for the channel, but at the cost of a large ($\operatorname{poly}(d)$) amount of nontrivial quantum computation and a copy complexity of $1/\epsilon^5$ with respect to the additive error. Our procedure uses much more trivial quantum resources, and with a circuit depth that is independent of the dimension. The idea of tomography of matrix product states [49] has been extended to matrix product operators that are unitary in the numerical work of Ref. [50]. There is as yet no systematic mathematical treatment of matrix product operator tomography, or tomography of quantum channels with bounded correlations of the kind discussed in this paper.

The above tomographic techniques only provide *additive* precision estimates rather than the *relative* precision estimates achieved here. Furthermore, these tomographic methods are not robust against SPAM errors, which create an unknown systematic error, or, equivalently, places an absolute floor on the achievable additive precision that can be obtained using these tomographic techniques [51]. The magnitude and complexity of SPAM errors scales with the system size, so that the achievable precision decreases for larger numbers of qubits. These tomographic techniques could be made robust to SPAM errors using gate-set tomography [51, 52], at the cost of substantially increasing the resources required. However, it is unclear how to obtain relative-precision estimates from any of these methods.

We achieve multiplicative precision and robustness to SPAM errors by using techniques originating from randomized benchmarking. Randomized benchmarking tomography [53] also employs randomized benchmarking techniques to achieve robustness to SPAM, and can obtain tomographic reconstructions using $O\big(d^2 \log(d)\big)$ randomized benchmarking experiments with an additive reconstruction error that scales as $O(d/\sqrt{\log d})$ [54]. Note that the additive error is essentially fundamental to the approach used in randomized benchmarking tomography because the decay rates are the fidelities between the noisy process and distinct Clifford gates. These fidelities are significantly less than 1—even for ideal noise processes—resulting in rapid decays that are difficult to estimate precisely.

Besides the prior work on quantum tomography, there has also been a large amount of prior work on learning Markov random fields, both in terms of parameter estimation (including hardness results [55]) and in terms of learning the structure of the factor graph [56–61]. These results are certainly relevant, but to make a direct comparison to the present results is complicated by the fact that our probability distribution contains variables that might be called "quasi-latent", in the sense that they are not observable to all queries, only to some, and queries that probe one variable must necessarily hide others. This is due to the symplectic structure of the Pauli group: we can only simultaneously measure observables that commute. Therefore learning a "symplectic Markov field" seems to be an inherently different task from what has been previously considered.

II. MATHEMATICAL PRELIMINARIES

Given a set of n qubits with Hilbert space dimension $d = 2^n$, let us introduce the following notation. Let $\mathbb{P}^n$ denote the group of Pauli operators acting on n qubits, and $\mathsf{P}^n = \mathbb{P}^n/\langle i \rangle$ be the quotient of $\mathbb{P}^n$ with its center. The group P^n is Abelian and isomorphic to $\mathbb{Z}_2^{2n}$, so we will label elements in P^n by bit strings of length $2n$. We will abuse notation and talk about $a \in \mathsf{P}^n$ when we mean a matrix $P_a \in \mathsf{P}^n$ up to an overall phase. For concreteness, for a bit string a and a single-qubit Pauli operator A we write $A[a] = \bigotimes_{a_i \in a} A^{a_i}$ and choose the elements of P^n to be

$$P_a = P_{(a_x, a_z)} = i^{a_x \cdot a_z} X[a_x] Z[a_z] \tag{1}$$

where $a \in \mathbb{Z}_2^{2n}$ and X and Z are the standard single-qubit Pauli matrices. The choice of phase is arbitrary but this choice of phase ensures that every P_a is Hermitian. For any two Pauli matrices P_a and P_b with respective images a and b in $\mathbb{Z}_2^{2n}$, we have $P_a P_b = (-1)^{\langle a,b \rangle} P_b P_a$ where

$$\langle a, b \rangle = a_x \cdot b_z + a_z \cdot b_x \mod 2 \tag{2}$$

is a binary symplectic form, and so it is symmetric and linear in each argument. We will typically omit the mod 2 because it is clear from context.

We define a stabilizer group to be a linear subspace of $\mathbb{Z}_2^{2n}$ such that $\langle a, b \rangle = 0$ for all $a, b \in \mathsf{S}$. Therefore every stabilizer group is a group $\mathsf{S} \subset \mathsf{P}^\mathsf{n}$ whose elements all commute as matrices. Note that stabilizer groups are often defined in terms of the matrices with specific phase conventions [62].

The commutant of a *set* $\mathsf{G} \subseteq \mathsf{P}^n$ is the group $\mathsf{C}_\mathsf{G} \subseteq \mathsf{P}^n$ that is orthogonal to G according to the binary symplectic form. That is,

$$\mathsf{C}_\mathsf{G} = \{a \in \mathsf{P}^n : \forall g \in \mathsf{G},\ \langle a, g \rangle = 0\}. \tag{3}$$

We refer to C_G as the commutant because every element of C_G commutes with every element of G as matrices in $\mathbb{P}^n$. When G is a group (i.e., closed under multiplication), $\mathsf{C}_{\mathsf{C}_\mathsf{G}} = \mathsf{G}$ by the double commutant theorem. We define the anti-commutant A_G of $\mathsf{G} \subseteq \mathsf{P}^n$ to be the quotient group $\mathsf{P}^n/\mathsf{C}_\mathsf{G}$. Any $b \in \mathsf{P}^n$ can be uniquely decomposed as

$$b = a + c \tag{4}$$

for some $a \in \mathsf{A}_\mathsf{G}$ and $c \in \mathsf{C}_\mathsf{G}$.

Any element of the Pauli group P^n commutes with exactly half of the other elements in the group. A similar statement is true for any subgroup of P^n, and we have the following lemma.

Lemma 1. *For any Pauli $a \in \mathsf{P}^n$ and any group $\mathsf{G} \subseteq \mathsf{P}^n$,*

$$\frac{1}{|\mathsf{G}|} \sum_{b \in \mathsf{G}} (-1)^{\langle a,b \rangle} = 1[a \in \mathsf{C}_\mathsf{G}].$$

Our Pauli estimation procedure varies over states and measurements whose stabilizer groups have elements that cover some set X of errors. This motivates the following definition.

Definition 1 (Stabilizer coverings)**.** A *stabilizer covering* O of a set $\mathsf{X} \subseteq \mathsf{P}^n$ is a set of stabilizer groups $\mathsf{O} = \{\mathsf{S}_j\}$ such that $\mathsf{X} \subseteq \bigcup_j \mathsf{S}_j$.

For important applications, X will not necessarily be a group. For example, X could be the set of Pauli errors with weight $\leq w$, which is not generally a group. The next lemma bounds the size of a stabilizer covering for any set X, and in some cases we can improve over the trivial bound.

Lemma 2. *For any set* $\mathsf{X} \in \mathsf{P}^n$ *there exists a stabilizer covering* O *of* X *with cardinality*

$$|\mathsf{O}| \leq \min\Big\{|\mathsf{X}|, \sqrt{|\langle \mathsf{X} \rangle / \mathsf{S}_{\mathsf{X}}|} + 1\Big\}, \tag{5}$$

where $\langle \mathsf{X} \rangle$ *is the group generated by the elements of* X *and* S_{X} *is any maximal stabilizer subgroup of* $\langle \mathsf{X} \rangle$*. In particular, the bound is tight when* X *is a group, and* $|\mathsf{O}| \leq 2^n + 1$ *unconditionally.*

Proof. For any set X, a simple construction of a stabilizer covering is the set of all two-element stabilizer groups generated by each nontrivial element of X. That is, $\mathsf{O} = \{\langle r \rangle : r \in \mathsf{X} \backslash \{0\}\}$. Clearly this set has $|\mathsf{O}| \leq |\mathsf{X}|$.

Now consider a stabilizer covering of $\langle \mathsf{X} \rangle$, which obviously will also cover X. As a subgroup of P^n, the generators can be partitioned into s generators of a stabilizer group S_{X} and $2k$ generators of the quotient $\langle \mathsf{X} \rangle / \mathsf{S}_{\mathsf{X}}$. The quotient comprises k logical qubits, so it is isomorphic to P^k and has 4^k elements. Then we can construct O to be the stabilizer groups of a complete set of $|\mathsf{O}| = 2^k + 1$ mutually unbiased bases (MUBs) [63] acting on the k logical qubits, where each MUB is taken in direct sum with S_{X}. The cardinality of this choice for O is just $2^k + 1 = \sqrt{4^k} + 1$, as claimed.

The MUBs give a minimal covering when X is a group, as there are at most $2^k - 1$ non-identity Pauli elements in a stabilizer group on k qubits, so at least $2^k + 1$ stabilizer groups are required to cover all $4^k - 1$ non-identity elements of P^k. □

A. States and measurements

We will use the following states and measurements constructed from stabilizer groups. For every $a \in \mathsf{P}^n$ and every stabilizer group S,

$$\rho_{\mathsf{S},a} = \frac{1}{d} \sum_{s \in \mathsf{S}} (-1)^{\langle s,a \rangle} P_s \tag{6}$$

is a valid quantum state, known as a stabilizer state. Note that for any $a \in \mathsf{P}^n$ and $b \in \mathsf{C}_{\mathsf{S}}$, $\rho_{\mathsf{S},a+b} = \rho_{\mathsf{S},a}$ and so we can uniquely label stabilizer states by the error syndromes $e \in \mathsf{A}_{\mathsf{S}}$. Moreover,

$$E_{\mathsf{S}} = \left\{ E_{\mathsf{S},e} = \frac{d}{|\mathsf{A}_{\mathsf{S}}|} \rho_{\mathsf{S},e} : e \in \mathsf{A}_{\mathsf{s}} \right\} \tag{7}$$

is a valid measurement, otherwise known as a syndrome measurement. The normalization is such that each $E_{\mathsf{S},e}$ is an orthogonal projector. If $\dim \mathsf{S} = n$, then $\rho_{\mathsf{S},e}$ is a pure state and E_{S} is a rank-1 projective measurement.

B. Quantum channels

We wish to partially characterize a general linear map $\mathcal{L} : \mathbb{C}^{d \times d} \to \mathbb{C}^{d \times d}$ which has a Kraus operator representation

$$\mathcal{L}(M) = \sum_k A_k M B_k^\dagger. \tag{8}$$

We will generally be interested in completely positive (CP) maps, where we can choose $A_k = B_k$ for all k. A specific class of channels that we frequently use are CP maps with a single unitary Kraus operator U. For any unitary matrix $U \in U(d)$, we implicitly define the corresponding ideal channel $\mathcal{U}$ that maps an arbitrary matrix M to $UMU^\dagger$. Note that the implicit function is many to one, as it maps phase multiples of a unitary matrix to the same channel. For any $a, b \in \mathsf{P}^{\mathsf{n}}$,

$$\mathcal{P}_a(P_b) = P_a P_b P_a^\dagger = (-1)^{\langle a,b \rangle} P_b. \tag{9}$$

We will partially characterize a general linear map by performing a Pauli twirl of the channel to reduce the effective channel to a Pauli channel, that is, to a linear map with a Kraus representation

$$\mathcal{E}(\rho) = \sum_{a \in \mathsf{P}^n} p_a P_a \rho P_a \tag{10}$$

where the p_a are *Pauli error rates* with $p_a \geq 0$ for CP maps. For any set $\mathbb{T} \subset U(d)$, we define the $\mathbb{T}$-twirl of a channel $\mathcal{L}$ to be

$$\mathcal{L}^{\mathbb{T}} = \frac{1}{|\mathbb{T}|} \sum_{T \in \mathbb{T}} \mathcal{T} \mathcal{L} \mathcal{T}^{\dagger}. \tag{11}$$

In Kraus form, we have

$$\mathcal{L}^{\mathbb{T}}(M) = \frac{1}{|\mathbb{T}|} \sum_{T \in \mathbb{T}} \sum_{k} \mathcal{T}(A_k) \rho \mathcal{T}(B_k)^{\dagger}. \tag{12}$$

As we now show, the Pauli-twirled channel $\mathcal{L}^{\mathsf{P}^n}$ is a Pauli channel where the error rates are directly related to the Kraus operators of the untwirled channel $\mathcal{L}$.

Lemma 3 (Pauli error rates)**.** *For a linear map $\mathcal{L}(M)$ as in eq. (8) with Kraus operators $A_k = \sum_{a \in \mathsf{P}^n} l_{k,a} P_a$ and $B_k = \sum_{b \in \mathsf{P}^n} r_{k,b} P_b$, the Pauli error rates p_a of the twirled channel $\mathcal{L}^{\mathsf{P}^n}$ are given by*

$$p_a = \sum_k l_{k,a} r^*_{k,a}. \tag{13}$$

Proof. From eq. (12),

$$\mathcal{L}^{\mathsf{P}^n}(M) = |\mathsf{P}^n|^{-1} \sum_{c \in \mathsf{P}^n} \sum_k \mathcal{P}_c(A_k) M \mathcal{P}_c(B_k)^{\dagger}.$$

As the Pauli matrices are a Hermitian orthogonal basis for $\mathbb{C}^{d \times d}$, we can expand the Kraus operators as $A_k = \sum_{a \in \mathsf{P}^n} l_{k,a} P_a$ and $B_k = \sum_{b \in \mathsf{P}^n} r_{k,b} P_b$, so that

$$\begin{aligned} \mathcal{L}^{\mathsf{P}^n}(M) &= |\mathsf{P}^n|^{-1} \sum_k \sum_{a,b,c \in \mathsf{P}^n} l_{k,a} r^*_{k,b} (-1)^{\langle c, a+b \rangle} P_a M P_b \\ &= \sum_k \sum_{a,b \in \mathsf{P}^n} l_{k,a} r^*_{k,b} 1[a=b] P_a M P_b \\ &= \sum_{a \in \mathsf{P}^n} \left(\sum_k l_{k,a} r^*_{k,a} \right) P_a M P_a, \end{aligned} \tag{14}$$

where in the second line we have used lemma 1 with the fact that $\mathsf{C}_{\mathsf{P}^n} = 0$. □

In the most important case of a CP map, $l_{k,a} = r_{k,a}$ and the error rates are manifestly nonnegative. However, effective negative error rates are possible in the presence of initial correlations, and this would correspond to the inner product in eq. (13) being negative. It is also noteworthy that this definition is independent of the freedom in the Kraus operators of a CP map, since the unitary freedom on the index k above does not change the inner product.

We can simplify the representation of Pauli channels using the vectorization map, a linear map $|\cdot) : X \to |X)$ that acts by stacking the columns of X in a given basis and satisfies the identity $|AXB^{\dagger}) = B^T \otimes A|X)$. In this notation, the dual vectors map to scalars via the Hilbert-Schmidt inner product, $(A|B) = \mathrm{Tr}(A^{\dagger}B)$. Then we can write the action of a channel $\mathcal{L}$ as a superoperator acting via left multiplication so that $\mathcal{L}|\rho) = \big|\mathcal{L}(\rho)\big)$. Note that we overload our notation by using the same symbol for the abstract channel and the superoperator matrix.

Any Pauli channel can be expressed as a superoperator in the Pauli basis as

$$\mathcal{E} = \frac{1}{d} \sum_{a,b \in \mathsf{P}^n} (-1)^{\langle a,b \rangle} p_b |P_a)(P_a|. \tag{15}$$

It is convenient to make the following two simplifications. First, we note that $|P_a)$ is not a normalized vector, so we adopt the convention that

$$|a) = \tfrac{1}{\sqrt{d}}|P_a) \tag{16}$$

so that $(a|b) = \frac{1}{d}\mathrm{Tr}(P_a^\dagger P_b) = \delta_{a,b}$. This notation is unambiguous as long as we distinguish carefully between a and P_a from the given context inside the vectorization map. Second, for any two sets of Pauli matrices $\mathsf{A}, \mathsf{B} \subseteq \mathsf{P}^n$, we define a Walsh-Hadamard transform $\boldsymbol{W}_{\mathsf{A},\mathsf{B}} : \mathbb{C}^{|\mathsf{B}|} \to \mathbb{C}^{|\mathsf{A}|}$ whose matrix representation is

$$\boldsymbol{W}_{\mathsf{A},\mathsf{B}} = \sum_{a\in\mathsf{A},b\in\mathsf{B}} (-1)^{\langle a,b\rangle}|a)(b|. \tag{17}$$

As we prove in lemma 4 below, this transformation is proportional to an isometry for specific groups A and B.

Given these definitions, the vector of Pauli error rates $\boldsymbol{p}$ is related to the superoperator representation in terms of the *Pauli fidelities*, defined as,

$$\boldsymbol{f} = \boldsymbol{W}\boldsymbol{p}, \tag{18}$$

where we have simplified the notation by using $\boldsymbol{W} = \boldsymbol{W}_{\mathsf{P},\mathsf{P}}$. After this change of variables we have simply

$$\mathcal{E} = \sum_{a\in\mathsf{P}^n} f_a|a)(a|. \tag{19}$$

With this convention, $f_0 = \sum_a p_a \leq 1$ with equality if the channel is trace preserving, and all other channel eigenvalues lie inside the interval $[-f_0, f_0]$ if the channel is completely positive. We are most interested in channels that are close to the identity in the sense that the eigenvalues are in some small, strictly positive interval $[\epsilon, 1]$, which is appropriate when the noise is weak. Consequently, for any $h \in \mathsf{P}^n$ and any set $\mathsf{A} \subseteq \mathsf{P}^n$ we define

$$r_h = 1 - f_h, \quad \boldsymbol{r}_{\mathsf{A}} = \sum_{a\in\mathsf{A}} r_a|a), \quad \boldsymbol{1}_{\mathsf{A}} = \sum_{a\in\mathsf{A}} |a). \tag{20}$$

We will omit the subscripts from the vector notation when $\mathsf{A} = \mathsf{P}^n$.

In what follows, our estimation strategy will be to infer both $\boldsymbol{r}$ and, by then doing the inverse transform, $\boldsymbol{p} = \boldsymbol{W}^{-1}(\boldsymbol{1} - \boldsymbol{r})$, where we can evaluate the inverse using the following lemma.

Lemma 4. *For any subgroups* $\mathsf{A}, \mathsf{B} \subseteq \mathsf{P}^n$, $\boldsymbol{W}_{\mathsf{A},\mathsf{B}}$ *satisfies*

$$\boldsymbol{W}_{\mathsf{A},\mathsf{B}}^\dagger \boldsymbol{W}_{\mathsf{A},\mathsf{B}} = |\mathsf{A}| \sum_{b,b'\in\mathsf{B}:b+b'\in\mathsf{C}_{\mathsf{A}}} |b')(b|.$$

In particular, when $\mathsf{C}_{\mathsf{A}} \cap \mathsf{B}$ *is trivial, we have that* $\boldsymbol{W}_{\mathsf{A},\mathsf{B}}$ *is proportional to an isometry, with* $\boldsymbol{W}_{\mathsf{A},\mathsf{B}}^\dagger \boldsymbol{W}_{\mathsf{A},\mathsf{B}} = |\mathsf{A}|\Pi_{\mathsf{B}}$ *where* $\Pi_{\mathsf{B}} = \sum_{b\in\mathsf{B}} |b)(b|$ *is the projector onto* B.

Proof. From eq. (17),

$$\begin{aligned}
\boldsymbol{W}_{\mathsf{A},\mathsf{B}}^\dagger \boldsymbol{W}_{\mathsf{A},\mathsf{B}} &= \sum_{a,a'\in\mathsf{A}} \sum_{b,b'\in\mathsf{B}} (-1)^{\langle a,b\rangle + \langle a',b'\rangle} |b')(a'|a)(b| \\
&= \sum_{b,b'\in\mathsf{B}} \left(\sum_{a\in\mathsf{A}} (-1)^{\langle a,b+b'\rangle} \right) |b')(b| \\
&= |\mathsf{A}| \sum_{b,b'\in\mathsf{B}} 1[b' + b \in \mathsf{C}_{\mathsf{A}}] |b')(b|,
\end{aligned}$$

where we have used lemma 1 to obtain the third line.

When B is a group and $\mathsf{C}_{\mathsf{A}} \cap \mathsf{B}$ is trivial, $b + b' \in \mathsf{C}_{\mathsf{A}}$ if and only if $b' = -b = b$. □

Finally, we can evaluate the quality of a reconstructed set of errors $\hat{\boldsymbol{p}}$ either in terms of a norm $\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|$ or a figure of merit on the space of linear maps. The natural norm would be the 1-norm, which is related to the average gate infidelity $r(\mathcal{E})$ of $\mathcal{E}$ to the identity channel and diamond distance from the identity $\epsilon_\diamond(\mathcal{E})$ by

$$\tfrac{1}{2}\||0) - \boldsymbol{p}\|_1 = \epsilon_\diamond(\mathcal{E}) = (1 + 1/d)r(\mathcal{E}), \tag{21}$$

which holds for Pauli channels [64].

However, motivated by physical considerations we will project reconstructions into the nearest point in the set of probability distributions, or potentially to the set of subnormalized probability distributions. The nearest point is not uniquely defined in general according to some norms, such as the 1-norm, and so in some cases we will instead use the 2-norm for which the nearest point is uniquely defined.

C. Model assumptions

Our procedure makes use of the following primitives: preparations of stabilizer states, Pauli transformations, and syndrome measurements. An experimental implementation of our procedure will necessarily involve noisy versions of all primitives. For clarity, we always denote the noisy version of a primitive with an overset $\sim$, e.g., $\tilde{A}$ is a noisy implementation of the ideal operation A (whether A is a state preparation, a measurement operator, or a channel).

For ease of analysis, we assume that the noise on the primitives is independent, time-stationary, and Markovian. In particular, we assume that the noisy preparation of a state ρ and a noisy syndrome measurement of a stabilizer group H are independent of the circuit to be applied, so that they can be written as a density operator $\tilde{\rho}$ and a positive-operator-valued measure (POVM) $\{\tilde{E}_{\mathsf{H},e} : e \in \mathsf{E}_{\mathsf{H}}\}$ respectively.

Similarly, we require that the noise in the implementations of a set of twirling channels $\mathcal{T}$ are independent of the particular twirl being implemented and the remainder of the circuit and so can be written as $\tilde{\mathcal{T}} = \mathcal{T}\Lambda$ for some fixed completely positive (CP) map Λ. We codify this in the following definition.

Definition 2 (GTM noise)**.** A noise model is time-stationary if the noisy implementation $\tilde{\mathcal{U}}(t)$ of a gate $\mathcal{U}$ at time t is a linear map that is independent of t and if state preparations and measurements are respectively described by fixed density operators and POVMs. A noise model for the Pauli group is called *GTM* (gate-independent, time-stationary, Markovian) if it is time-stationary and there exists a completely positive trace-preserving map Λ such that $\tilde{\mathcal{P}} = \mathcal{P}\Lambda$ for all $\mathcal{P} \in \mathsf{P}^n$.

The above assumption is routinely assumed in analyses of RB for groups X that contain complex circuits of multi-qubit gates and can be relaxed with sufficient effort [22–24]. In contrast, we only make the assumption for groups consisting of tensor products of channels acting on individual qubits.

The other assumption we will make on the noise is that it is sufficiently weak that we can ignore certain algebraic limitations in the use of our specific estimators. In particular, we will assume that our noise and the state preparation and measurements (SPAM) satisfy the following definitions for choices of a parameter c that we will specify later.

Definition 3 (Weak, stable)**.** A noise map $\mathcal{L}$ is ***c*-weak** if the Pauli twirl $\mathcal{L}^{\mathsf{P}^n}$ is close to the identity channel in the operator norm, $\|\mathcal{I} - \mathcal{L}^{\mathsf{P}^n}\| \leq c$. A SPAM parameter A is called ***c*-stable** if $A \geq 1 - c$.

The first definition is equivalent to saying that the Pauli twirl $\mathcal{L}^{\mathsf{P}^n}$ has Pauli fidelities $\boldsymbol{f}$ that all lie in the interval $[1 - c, 1]$. It is likely that this assumption too can be relaxed, but to do so would require different estimators than the ratio estimator that we use below. We note that a simple sufficient condition to ensure that a CP map is c-weak is that the Pauli error rate for the identity, p_0, obeys $p_0 \geq 1 - \frac{c}{2}$. This can be verified from the inequality $f_i \geq 2p_0 - 1$ [21]. The second definition will become clearer once we define (in eq. (23) below) how the SPAM parameters A_j in the definition depend on the noisy state preparations and measurements. Intuitively, it just ensures that the measurements are sufficiently good to give a reasonable signal, so a c-stable SPAM parameter is c-close to being ideal.

III. ESTIMATION PROCEDURE AND SAMPLE COMPLEXITY

We now specify a procedure for estimating the Pauli fidelities of a noisy implementation of the Pauli group P^n, where we prepare states using a fixed stabilizer group G and we make syndrome measurements of a second, possibly different, stabilizer group H. As we will see below, for this procedure to yield useful information about the noise we will choose $\mathsf{H} \subseteq \mathsf{G}$, but we do not require this choice in our proofs. The backbone of the procedure is running a generalized cycle benchmarking sequence of length m [21] and recording the result of a syndrome measurement. One of the core innovations of the present procedure is to consider measurements of general stabilizer groups, which enables us to obtain more information from each experiment. Character benchmarking [20] over the Pauli group, direct RB [65] with the Pauli group as the generator distribution, and cycle benchmarking with no interleaved gate [21] can all be rephrased in terms of the following more general procedure by choosing the measurement to be of a two-element stabilizer group, that is, a group generated by a single Pauli operator.

RunCB$(\mathsf{G}, \mathsf{H}, m)$

1. Prepare the approximate stabilizer state $\tilde{\rho}_{\mathsf{G},0}$.
2. For each $i = 0, \dots, m$, apply a Pauli gate $P_{a_i} \in \mathsf{P}^n$ uniformly at random.
3. Perform a syndrome measurement of H and record the outcome $b \in \mathsf{A_H}$ from the noisy POVM element $\tilde{E}_b$.
4. Return $z \in \mathsf{A_H}$ such that $b + \sum_i a_i \equiv z$.

We will use the shorthand **RunCB**$(\mathsf{G}, m) =$ **RunCB**$(\mathsf{G}, \mathsf{G}, m)$ for the special case $\mathsf{G} = \mathsf{H}$. The following proposition exactly characterizes the output of **RunCB**$(\mathsf{G}, \mathsf{H}, m)$ under gate-independent and time-stationary noise. We note that **RunCB**$(\mathsf{G}, \mathsf{H}, m)$ can be straightforwardly generalized to include interleaved gates [21]; this will be explored in future work.

Proposition 5. *Let $Z = Z(\mathsf{G}, \mathsf{H}, m)$ be the random variable taking values $z \in \mathsf{E_H}$ that results from one call to the subroutine* **RunCB**$(\mathsf{G}, \mathsf{H}, m)$. *Under GTM noise with $\tilde{\mathcal{P}} = \mathcal{P}\Lambda$ for all $P \in \mathsf{P}^n$, Z has the probability distribution function*

$$\Pr\big(Z = z\big) = \frac{1}{|\mathsf{A_H}|} \sum_{h \in \mathsf{H}} f_h^m (-1)^{\langle h, z \rangle} A_h \tag{22}$$

where the $f_h = (h|\Lambda|h)$ are the Pauli fidelities of Λ, and

$$A_h = \sum_{b \in \mathsf{A_H}} (-1)^{\langle h, b \rangle} (\tilde{E}_{\mathsf{H},b}|h)(h|\Lambda|\tilde{\rho}_{\mathsf{G},0}). \tag{23}$$

Proof. By the linearity of the Born rule for density matrices, the probability of a given instantiation $Z = z$ is the sum of all the probabilities of random noisy Pauli gates and measurement outcomes that would lead to that result. This is equivalent to first writing the entire joint distribution over any possible sequence, and then marginalizing over the specific sequences of Pauli gates and measurement outcomes that give the same value of z.

Under the assumption of GTM noise, the probability of choosing $a_0, a_1, \dots, a_m$ and obtaining the outcome b is

$$\Pr(a_0, \dots, a_m, b) = |\mathsf{P}^n|^{-m} (\tilde{E}_{\mathsf{H},b}| \left(\prod_{i=m}^{1} \mathcal{P}_{a_i} \Lambda \right) |\tilde{\rho}_{\mathsf{G},0}),$$

where for non-commutative products we use the convention $\prod_{i=a}^{b} x_i = x_a \dots x_b$ (i.e., the lower index is on the left of the product). We can rewrite the probability as

$$\Pr(a_0, \dots, a_m, b) = |\mathsf{P}^n|^{-m} (\tilde{E}_{\mathsf{H},b}|\mathcal{P}_{a'_m} \left(\prod_{i=m-1}^{0} \mathcal{P}_{a'_i} \Lambda \mathcal{P}_{a'_i} \right) \Lambda|\tilde{\rho}_{\mathsf{G},0}),$$

where we set $a'_0 = a_0$ and recursively define $a'_i = a_i + a'_{i-1}$ for $i = 1, \dots, m-1$, where $a'_m = \sum_i a_i$ (recall that addition is mod 2). Averaging uniformly and independently over the a_i is equivalent to averaging uniformly and independently over the a'_i because P^n is a group. Therefore the marginal probability of $a'_m \in \mathsf{P}^n$ and $b \in \mathsf{E}_\mathsf{H}$ is

$$\begin{aligned}
\Pr(a'_m, b) &= |\mathsf{P}^n|^{-m} \sum_{a'_1, \dots, a'_{m-1} \in \mathsf{P}^n} (\tilde{E}_{\mathsf{H},b}|\mathcal{P}_{a'_m} \left(\prod_{i=m-1}^{0} \mathcal{P}_{a'_i} \Lambda \mathcal{P}_{a'_i} \right) \Lambda|\tilde{\rho}_{\mathsf{G},0}) \\
&= |\mathsf{P}^n|^{-1} (\tilde{E}_{\mathsf{H},b}|\mathcal{P}_{a'_m} \left(\prod_{i=m-1}^{0} \Lambda^{\mathsf{P}^n} \right) \Lambda|\tilde{\rho}_{\mathsf{G},0}) \\
&= |\mathsf{P}^n|^{-1} \sum_{h \in \mathsf{P}^n} f_h^m (\tilde{E}_{\mathsf{H},b}|\mathcal{P}_{a'_m}|h)(h|\Lambda|\tilde{\rho}_{\mathsf{G},0}) \\
&= |\mathsf{P}^n|^{-1} \sum_{h \in \mathsf{P}^n} (-1)^{\langle h, a'_m \rangle} f_h^m A_{b,h}
\end{aligned}$$

by eqs. (9) and (19), where in particular we use $(h|\Lambda^{\mathsf{P}^n}|h) = (h|\Lambda|h) = f_h$ and we define $A_{b,h} = (\tilde{E}_{\mathsf{H},b}|h)(h|\Lambda|\tilde{\rho}_{\mathsf{G},0})$.

Splitting $a'_m = a + c$ where $a \in \mathsf{A}_\mathsf{H}$ and $c \in \mathsf{C}_\mathsf{H}$ and averaging over c gives the marginal probability

$$\begin{aligned}
\Pr(a, b) &= \sum_{c \in \mathsf{C}_\mathsf{H}} \Pr(a + c, b) \\
&= |\mathsf{A}_\mathsf{H}|^{-1} \sum_{h \in \mathsf{P}^n} (-1)^{\langle h, a \rangle} f_h^m A_{b,h} |\mathsf{C}_\mathsf{H}|^{-1} \sum_{c \in \mathsf{C}_\mathsf{H}} (-1)^{\langle h, c \rangle} \\
&= |\mathsf{A}_\mathsf{H}|^{-1} \sum_{h \in \mathsf{P}^n} (-1)^{\langle h, a \rangle} f_h^m A_{b,h} \, 1\big[h \in \mathsf{C}_{\mathsf{C}_\mathsf{H}}\big]
\end{aligned}$$

by lemma 1. Now by the double commutant theorem we have $\mathsf{C}_{\mathsf{C}_\mathsf{H}} = \mathsf{H}$ for any group $\mathsf{H} \subseteq \mathsf{P}^n$, so

$$\Pr(a, b) = |\mathsf{A}_\mathsf{H}|^{-1} \sum_{h \in \mathsf{H}} (-1)^{\langle h, a \rangle} f_h^m A_{b,h} \,.$$

The probability of obtaining the outcome $z \in \mathsf{A}_\mathsf{H}$ is then

$$\begin{aligned}
\Pr(Z = z) &= \sum_{a, b \in \mathsf{A}_\mathsf{H} : a + b = z} \Pr(a, b) \\
&= \sum_{b \in \mathsf{A}_\mathsf{H}} \Pr(z + b, b) \\
&= |\mathsf{A}_\mathsf{H}|^{-1} \sum_{b \in \mathsf{A}_\mathsf{H}} \sum_{h \in \mathsf{H}} (-1)^{\langle h, b + z \rangle} f_h^m A_{b,h} \\
&= |\mathsf{A}_\mathsf{H}|^{-1} \sum_{h \in \mathsf{H}} (-1)^{\langle h, z \rangle} f_h^m A_h
\end{aligned}$$

which completes the proof. □

We remark that the coefficients A_h appearing in proposition 5 are 1 for ideal states, measurements, and transformations. Indeed, setting $\tilde{\rho}_{\mathsf{G},0} = \rho_{\mathsf{G},0}$, $\Lambda = 1$ and $\tilde{E}_{\mathsf{H},b} = E_{\mathsf{H},b}$ for any fixed outcome $b \in \mathsf{A_H}$ in eq. (23) and substituting in eqs. (6) and (7), we have that

$$\begin{aligned}(E_{\mathsf{H},b}|h)(h|\rho_{\mathsf{G},0}) &= \frac{1}{d|\mathsf{A_H}|}\sum_{g\in\mathsf{G}}\sum_{h'\in\mathsf{H}}(-1)^{\langle h',b\rangle}(P_{h'}|h)(h|P_g)\\ &= \frac{(-1)^{\langle h,b\rangle}}{|\mathsf{A_H}|}1[h \in \mathsf{G}\cap\mathsf{H}],\end{aligned}$$

where we have used the normalization from eq. (16). Plugging this into eq. (23), we find that in the ideal case,

$$A_h = 1[h \in \mathsf{G}\cap\mathsf{H}]. \qquad \text{(ideal case)}$$

We refer to the coefficients A_h as SPAM coefficients (despite having a contribution from edge noise terms) because the coefficients are the only model parameters affected by errors in the state preparations and measurements. We see that the information contained in the likelihood function is maximized when $\mathsf{H} = \mathsf{G}$.

Proposition 5 exactly quantifies the output of one call to $\mathbf{RunCB}(\mathsf{G},\mathsf{H},m)$. One could naively obtain many samples of $\mathbf{RunCB}(\mathsf{G},\mathsf{H},m)$ for different values of m and perform a multi-exponential fit to estimate the model parameters $\{A_h, f_h : h \in \mathsf{H}\}$. However, there are exponentially many such parameters, making the fit and a rigorous theoretical treatment nigh impossible. We can resolve this problem by transforming the output using the following lemma to obtain a new random variable with a single exponential decay as in [20]. We note that considering syndrome measurements instead of the two-outcome measurements considered in Refs. [20, 21] allows us to estimate multiple random variables from the same data using the following procedure, which can significantly reduce the sample complexity.

```
$\mathbf{V}(\mathsf{X},\mathsf{G},t,m)$: Estimator function for vector of SPAM-dependent Pauli fidelities.
  1. Set $\boldsymbol{V} := 0$.
  2. For $k = 1,\ldots,t$, Do
       • Set $z := \mathbf{RunCB}(\mathsf{G},m)$,
       • Set $\boldsymbol{V}$ += $\sum_{x\in\mathsf{X}}(-1)^{\langle x,z\rangle}|x)$.
  3. Return $\hat{\boldsymbol{V}} = \frac{1}{t}\boldsymbol{V}$.
```

The elements in the output of $\mathbf{V}(\mathsf{X},\mathsf{G},t,m)$ are correlated binomial variables, but their covariance can also be exactly computed using the following lemma.

Lemma 6. *Let* $\boldsymbol{V}$ *be the output of* $\mathbf{V}(\mathsf{X},\mathsf{G},t,m)$. *Under GTM noise, we have*

$$\begin{aligned}\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{V}] &= \sum_{x\in\mathsf{X}} A_{x^\perp} f_{x^\perp}^m |x)\\ \mathbb{E}\big[\boldsymbol{V}\boldsymbol{V}^\dagger\big] &= \sum_{x,x'\in\mathsf{X}} 1[x + x' \in \mathsf{P^n/G}]\, \mathsf{A_{x+x'}}\mathsf{f^m_{x+x'}}|\mathsf{x})(\mathsf{x'}|\end{aligned}$$

where $x^\perp$ *is the component of* x *in* P^n/G *and the* A_x *are as in proposition 5, and, in particular, are independent of* m.

Proof. As the calls to $\mathbf{RunCB}(\mathsf{G},m)$ are independent, the expectation value of $\boldsymbol{V}$ is

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{V}] = \sum_{x\in\mathsf{X}}\sum_{z\in\mathsf{A_G}}(-1)^{\langle x,z\rangle}\Pr(Z = z)|x).$$

For any fixed $x \in \mathsf{X}$, substituting in proposition 5 with $\mathsf{H} = \mathsf{G}$, noting we $\mathbf{RunCB}(\mathsf{G},m)$), and using lemma 1 gives

$$\begin{aligned}(x|\boldsymbol{\mu} &= \frac{1}{|\mathsf{A_G}|}\sum_{z\in\mathsf{A_G}}\sum_{g\in\mathsf{G}}(-1)^{\langle x+g,z\rangle}f_g^m A_g\\ &= \sum_{g\in\mathsf{G}} 1[x + g \in \mathsf{C_{A_G}}]f_g^m A_g.\end{aligned}$$

Note that for a stabilizer group G, $\mathsf{C}_{\mathsf{A}_\mathsf{G}} = \mathsf{P}^n/\mathsf{G}$. The result then follows as G is a group and P^n splits into $\mathsf{G} \oplus \mathsf{P}^n/\mathsf{G}$.

Similarly, for fixed $x, x' \in \mathsf{X}$ we have

$$(x|\mathbb{E}\big[\boldsymbol{V}\boldsymbol{V}^\dagger\big]|x') = \sum_{z \in \mathsf{A}_\mathsf{H}} (-1)^{\langle x+x', z\rangle} \Pr(Z = z) = (x + x'|\mathbb{E}[\boldsymbol{V}],$$

which completes the proof. □

Since the elements of $\boldsymbol{V}$ are bounded, an average of sufficiently many independent samples will converge quickly to the mean. The next proposition provides a simple and conservative tail bound on the probability of estimates obtained using T samples will deviate from their mean by more than ϵ. Specifically, we bound the failure probability of the following procedure, where we allow $\mathsf{X} \subseteq \mathsf{G}$ both for later analysis and to illustrate the fundamental scaling.

Proposition 7. *Let* G *be a stabilizer group,* $\mathsf{X} \subseteq \mathsf{G}$*, and* $\hat{\boldsymbol{V}}$ *be the output of* $\boldsymbol{V}(\mathsf{X}, \mathsf{G}, t, m)$ *for some fixed positive integers* t *and* m*. Then for any* $\epsilon > 0$*,*

$$\Pr(\|\hat{\boldsymbol{V}} - \mathbb{E}(\hat{\boldsymbol{V}})\|_\infty \geq \epsilon) \leq 2|\mathsf{X}| \exp(-t\epsilon^2/2).$$

Proof. The proof follows trivially from the union bound and Hoeffding's inequality [66] where each element of $\hat{\boldsymbol{V}}$ is in the interval $[-1, 1]$. □

Note that the above bound is very loose because, as shown in lemma 6, there are correlations between the elements of $\hat{\boldsymbol{V}}$ that are neglected when applying the union bound. However, we still see that the sample complexity above in estimating $A_j f_j^m$ is independent of the Pauli fidelities. The precision in estimating the fidelities f_j with a fixed failure probability will, however, depend on the bare precision ϵ and on the specific values of m that we choose. In the next section we will analyze a specific procedure and estimator to get bounds on this dependency.

IV. DATA FITTING AND ERROR ANALYSIS

To understand how the sequence lengths should be chosen to get a precise estimate of a Pauli channel, we must introduce a specific estimation strategy. We use an estimator that samples Af^m at exponentially increasing values of m, and then we use only the data at the endpoints to obtain a ratio estimator. We also ignore repeated elements from distinct stabilizer groups in a stabilizer covering of a set. Our estimation strategy throws away a lot of data, but it has two advantages. First, it is easy to analyze explicitly, and our proof is a straightforward adaption of the theorem presented in Ref. [36]. Second, the endpoints are where the data are most sensitive, so discarding the intermediate data is not as damaging as one might expect since not all points contribute equally to the variance in the estimator. In what follows we use the notation $V_i(g) = (g|V_i$ for the gth element of V_i since the subscript real estate is occupied by the iteration index i.

```
Ratio$(\mathsf{O}, \mathsf{X}, t)$ Ratio estimator for exponential regression of a stabilizer covering $\mathsf{O}$ of a set $\mathsf{X}$.
  1. Set $\hat{r}_x = \texttt{NaN}$ for all $x \in \mathsf{X}$.
  2. Set $\mathsf{A} := \emptyset$.
  3. For all $\mathsf{G} \in \mathsf{O}$, Do
     (a) Set $\mathsf{X}' := \mathsf{G} \backslash \mathsf{A}$.
     (b) Set $\mathsf{A} := \mathsf{A} \cup \mathsf{G}$.
     (c) Set $\hat{V} := \mathbf{V}(\mathsf{X}', \mathsf{G}, t, 0)$.
     (d) Set $m := 1$.
     (e) While $\exists x \in \mathsf{X}'$ such that $\hat{r}_x = \texttt{NaN}$, Do
```

```
        • Set $\hat{W} := \mathbf{V}(\mathsf{X}', \mathsf{G}, t, m)$,
        • For all $x \in \mathsf{X}'$ such that $\hat{r}_x = \texttt{NaN}$, Do
          – Set $v := \hat{V}(x)$ and $w := \hat{W}(x)$.
          – If $w \leq v/3$ and $w, v > 0$, Set $\hat{r}_x := 1 - \big(w/v\big)^{1/m}$;
          – Else If $w \leq 0$ or $v \leq 0$, Set $\hat{r}_x := 1$.
        • Set $m := 2m$
4. Return $\hat{\boldsymbol{r}}_{\mathsf{X}} = \sum_{x \in \mathsf{X}} \hat{r}_x |x)$.
```

This ratio estimator guarantees a pointwise multiplicative precision estimate of $\boldsymbol{r}_{\mathsf{X}}$, at least when certain mild assumptions hold that place us close to the regime of interest. This is where the notions of c-weak and c-stable from section II C come into play. We will assume that the noise map is $\frac{1}{2}$-weak, which in particular implies that $f_h \geq \frac{1}{2}$, or equivalently, that $r_h \leq \frac{1}{2}$. We will also assume that each SPAM parameter A_h from eq. (23) is $\frac{1}{2}$-stable, so that $A_h \geq \frac{1}{2}$. However, the only thing that our proof really requires is that both A and f are bounded from below by a positive constant, and the specific choice of $\frac{1}{2}$ is motivated only by convenience.

We note that an equivalent theorem to the following can be proven if we remove the $\frac{1}{2}$-stable assumption and instead assume that ϵ is chosen to give a relative precision proportional to $\min_h A_h f_h$ at each step instead of only a constant. However, this makes the statements about sample complexity dependent on the SPAM parameters A_h. This obfuscates the actual sample complexity of the procedure, so we prefer to add the stability assumption.

To understand how the set of sequence lengths κ should be chosen to get a precise estimate of a Pauli channel, we must introduce the spectral gap of the channel, or of a subset of the channel corresponding to the Paulis in the set X. Since a Pauli channel is already diagonal in the Pauli basis, as in Eq. (19), and every trace-preserving channel has a largest absolute eigenvalue of 1, the spectral gap $\Delta_{\mathsf{X}} = \Delta_{\mathsf{X}}(\Lambda)$ of the Pauli channel Λ over the set X is given by

$$\Delta_{\mathsf{X}} = 1 - \max_{j \in \mathsf{X} \setminus 0} |f_j|. \tag{24}$$

In the most interesting regime, Λ is ϵ-weak for some $\epsilon < 1$, and the absolute value is not necessary because then all fidelities are positive. In that case, we have $\Delta_{\mathsf{X}} = \min_{x \in \mathsf{X}} r_x$. When $\mathsf{X} = \mathsf{P}^n$, this is exactly the spectral gap of the superoperator for the channel. Note that $\Delta_{\mathsf{X}} \geq \Delta_{\mathsf{P}^n}$.

As we will now show, one can completely learn all the Pauli fidelities in X with relative precision ϵ using a set κ of sequence lengths with $|\kappa| = O\big(\log \frac{1}{\Delta_{\mathsf{X}}}\big)$ and $m_{\max} = \max \kappa = O\big(\frac{1}{\Delta_{\mathsf{X}}}\big)$.

Proposition 8. *Let $\mathsf{X} \subseteq \mathsf{P}^n$ and O be a stabilizer covering of X. For any sufficiently small $\epsilon, \delta > 0$ and assuming $\frac{1}{2}$-weak, $\frac{1}{2}$-stable, GTM noise, the following holds with probability $1 - \delta$. Running* **Ratio**$(\mathsf{O}, \mathsf{X}, t)$ *with $t = \frac{2}{\epsilon^2} \log\big(\frac{2|\mathsf{X}||\kappa|}{\delta}\big)$ uses a set κ of at most $O\big(\log \frac{1}{\Delta_{\mathsf{X}}}\big)$ sequence lengths m with $m_{\max} = O\big(\frac{1}{\Delta_{\mathsf{X}}}\big)$. Moreover, the output $\hat{\boldsymbol{r}}_{\mathsf{X}}$ satisfies*

$$|\hat{r}_x - r_x| \leq O(\epsilon) r_x\,.$$

Proof. The following proof of multiplicative precision is a modification of a similar result proven recently in Ref. [36].

We begin by considering each iteration of the loop in step 3. Let t be a fixed integer, $\mathsf{G} \in \mathsf{O}$ be a fixed group, and X' be the value in step 3a in the corresponding iteration of **Ratio**$(\mathsf{O}, \mathsf{X}, t)$. The inner loop (e) yields estimators $\hat{\boldsymbol{V}}_m = \mathbf{V}(\mathsf{X}', \mathsf{G}, t, m)$ for each $m \in \kappa = \{0, 1, 2, 4, \ldots, m_{\max}\}$ for some yet-unspecified terminating sequence length $m_{\max}$. Note that we do not store all these intermediate estimators in the procedure as we only use them to prove correctness.

By proposition 7 and the union bound, for any fixed $\epsilon > 0$,

$$|\hat{V}_m(x) - A_x f_x^m| \leq \epsilon$$

with probability at least $1 - \delta_{\mathsf{X}'}$ where $\delta_{\mathsf{X}'} = 2|\kappa||\mathsf{X}'| \exp(-t\epsilon^2/2)$. To bound the terminating value $m_{\max}$ for each $\mathsf{G} \in \mathsf{O}$ and hence prove the stated sample complexity, fix $g \in \mathsf{G}$, let m

be the corresponding value in the procedure when $\hat{r}_x$ is assigned and $w = \hat{V}_m(\mathsf{X}', \mathsf{G}, t, m)$ and $u = \hat{V}_m(\mathsf{X}', \mathsf{G}, t, \lfloor m/2 \rfloor)$. Then u, v, w satisfy the following inequalities

$$\frac{w}{v} \leq \frac{1}{3} < \frac{u}{v}. \tag{25}$$

For simplicity, we now use the notation $f = f_g$ and $A = A_g$. We also assume that $m > 1$ as otherwise the proof is trivial. Let us define the deviations of u, v, w from their mean,

$$w - Af^m =: \epsilon_w\,, \quad u - Af^{m/2} =: \epsilon_u\,, \quad \text{and } v - A =: \epsilon_v\,.$$

By proposition 7, $|\epsilon_x| \leq \epsilon$ for $x = u, v, w$. Using the fact that $A \geq 1/2$ for $\frac{1}{2}$-stable noise,

$$\frac{w}{v} = \frac{Af^m + \epsilon_w}{A + \epsilon_v} = \left(1 + \tfrac{\epsilon_v}{A}\right)^{-1}\left(f^m + \tfrac{\epsilon_w}{A}\right) \geq (1 + 2\epsilon)^{-1}\left(f^m - 2\epsilon\right). \tag{26}$$

Similarly,

$$\frac{u}{v} \leq (1 - 2\epsilon)^{-1}\left(f^{m/2} + 2\epsilon\right) \tag{27}$$

$$\frac{w}{v} \leq (1 - 2\epsilon)^{-1}\left(f^m + 2\epsilon\right). \tag{28}$$

Substituting eqs. (26) and (27) into eq. (25), we have

$$\left(\frac{1 - 8\epsilon}{3}\right)^2 < f^m \leq \frac{1 + 8\epsilon}{3}. \tag{29}$$

We have that $r = 1 - f \geq \Delta_{\mathsf{X}}$, and note that $f^m = (1 - r)^m$ will satisfy eq. (29) for fixed ϵ if and only if $m = \Theta(1/r)$.

To show that $|\hat{r} - r| \leq O(\epsilon) r$, we have to analyze the accuracy of the estimator $(w/v)^{1/m}$. Using eqs. (26) and (28), we have

$$\left(\frac{f^m + 2\epsilon}{1 - 2\epsilon}\right)^{1/m} \geq \hat{f} \geq \left(\frac{f^m - 2\epsilon}{1 + 2\epsilon}\right)^{1/m}. \tag{30}$$

We can factor out the f^m (since $f > 0$ as we have assumed the noise is 1/2-weak) and use the lower bound from eq. (29) to find

$$f\left(\frac{1 - 18\epsilon/(1 - 8\epsilon)^2}{1 + 2\epsilon}\right)^{1/m} < \hat{f} < f\left(\frac{1 + 18\epsilon/(1 - 8\epsilon)^2}{1 - 2\epsilon}\right)^{1/m}. \tag{31}$$

For sufficiently small ϵ, this implies

$$f\big(1 - O(\epsilon)\big)^{1/m} \leq \hat{f} \leq f\big(1 + O(\epsilon)\big)^{1/m}. \tag{32}$$

Now we Taylor expand around $\epsilon = 0$ for sufficiently small ϵ and use $\frac{1}{m} = \Theta(r)$ to obtain the stated accuracy.

Now let $g \in \mathsf{X}$ be the element that requires the largest value of $m_{\max}$. Then the set of sequence lengths κ used to estimate all the f_a to within the stated precision with probability at least $1 - \delta_{\mathsf{X}'}$ is $\kappa = \{2^i : i = 0, \ldots, \ell = \log_2(m_{\max})\}$. As $m_{\max} = \Theta(1/r_g)$, we have $|\kappa| - 1 = \log_2 m_{\max} = O\big(\log \frac{1}{\Delta_{\mathsf{X}}}\big)$, which is well defined as the noise is 1/2-weak by assumption. The total failure probability is then at most δ where $\delta = 2|\kappa||\mathsf{X}|\exp(-2t\epsilon^2/2)$ by the union bound. Rearranging gives the stated sample complexity. □

V. RECONSTRUCTING ERROR RATES FOR A GROUP

We have seen how we can estimate each of the Pauli fidelities f_a of a Pauli channel for all $a \in \mathsf{X}$ with a number of measurements that scales like $O\big(|\mathsf{O}| \log |\mathsf{X}|\big)$ for any stabilizer covering O of X, assuming that ϵ, δ and the channel spectral gap Δ_{X} are fixed. When X is a group, we can choose O such that $|\mathsf{O}| \leq \sqrt{|\mathsf{X}|} + 1$ using the construction in lemma 2, so the total number of samples is at most $O(\sqrt{|\mathsf{X}|}\log|\mathsf{X}|)$ (or even less if X has a nontrivial stabilizer subgroup). Moreover, the precision of these estimates is multiplicative. We now show how the estimates of Pauli fidelities can be used to estimate Pauli error rates.

Suppose we estimate all the Pauli fidelities of some group G (which could be the full Pauli group). Then eq. (18) becomes

$$\boldsymbol{f}_{\mathsf{G}} = \boldsymbol{W}_{\mathsf{G},\mathsf{P}^n}\boldsymbol{p}, \tag{33}$$

where we use the shorthand $\boldsymbol{v}_{\mathsf{G}}$ to denote the vector with entries $\{v_g : g \in \mathsf{G}\}$. The columns of $\boldsymbol{W}_{\mathsf{G},\mathsf{P}^n}$ for Paulis that differ by an element of C_{G} are identical and so the corresponding Pauli error rates cannot be distinguished using only the estimated Pauli fidelities. That is, we can only reconstruct a *marginal* or coarse-grained probability distribution $\boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}$ over A_{G} with marginal probabilities

$$p_{\mathsf{A}_{\mathsf{G}},a} = \sum_{c\in\mathsf{C}_{\mathsf{G}}} p_{a+c} \tag{34}$$

for $a \in \mathsf{A}_{\mathsf{G}}$ via

$$\boldsymbol{f}_{\mathsf{G}} = \boldsymbol{W}_{\mathsf{G},\mathsf{A}_{\mathsf{G}}}\boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}. \tag{35}$$

When G is the group generated by the stabilizers and logical operators of an error-correcting code, the marginal probabilities give exactly the distribution of logical errors for each syndrome. This marginal distribution can therefore be used to construct the maximum likelihood decoder for the actual noise afflicting a device.

In practice, inverting eq. (35) to find the marginal error rates might be solved using, for example, a non-negative least squares solver, being careful to take the nontrivial covariance into account as well when estimating the errors. However, to prove concrete theorems, we will use the estimator

$$\hat{\boldsymbol{p}}_{\mathsf{A}_{\mathsf{G}}} = \left[\boldsymbol{W}^{-1}_{\mathsf{G},\mathsf{A}_{\mathsf{G}}}\hat{\boldsymbol{f}}_{\mathsf{G}}\right]_{\mathcal{D}} = \left[|\mathsf{G}|^{-1}\boldsymbol{W}_{\mathsf{A}_{\mathsf{G}},\mathsf{G}}\hat{\boldsymbol{f}}_{\mathsf{G}}\right]_{\mathcal{D}} \tag{36}$$

using lemma 4. Here we use the notation $[\boldsymbol{v}]_{\mathcal{D}}$ to mean taking a k-dimensional vector $\boldsymbol{v}$ and projecting it to the nearest point (according to a Euclidean metric) in the k-point simplex, which we denote $\mathcal{D}$, with the dimension being inferred from the context. The informal statement in result 1 follows as $\|\boldsymbol{r}_{\mathsf{G}}\|_\infty \leq 2(1-p_0)$ by [21, Lemma 4].

Proposition 9. *Let G be a group of Pauli matrices with a maximal stabilizer subgroup S. Then, under the conditions of proposition 8, an estimator $\hat{\boldsymbol{p}}_{\mathsf{A}_{\mathsf{G}}}$ of $\boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}$ satisfying*

$$\|\hat{\boldsymbol{p}}_{\mathsf{A}_{\mathsf{G}}} - \boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}\|_2 \leq O(\epsilon)\|\boldsymbol{r}_{\mathsf{G}}\|_\infty$$

with probability at least $1-\delta$ can be obtained using $t = \frac{2}{\epsilon^2}\log\left(\frac{2|\kappa||\mathsf{G}|}{\delta}\right)$ measurements per round for $|\kappa|(\sqrt{|\mathsf{G}/\mathsf{S}|}+1)$ rounds, where $|\kappa| = O\left(\log\frac{1}{\Delta_{\mathsf{G}}}\right)$ and $m_{\max} = O\left(\frac{1}{\Delta_{\mathsf{G}}}\right)$ is the longest sequence length used.

Proof. Suppose we have an estimator $\hat{\boldsymbol{r}}_{\mathsf{G}}$ satisfying

$$\|\hat{\boldsymbol{r}}_{\mathsf{G}} - \boldsymbol{r}_{\mathsf{G}}\|_\infty \leq \epsilon_0.$$

First note that projecting to the simplex can only decrease the error, that is,

$$\|\hat{\boldsymbol{p}}_{\mathsf{A}_{\mathsf{G}}} - \boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}\|_2 \leq \||\mathsf{G}|^{-1}\boldsymbol{W}_{\mathsf{A}_{\mathsf{G}},\mathsf{G}}(\hat{\boldsymbol{r}}_{\mathsf{G}} - \boldsymbol{r}_{\mathsf{G}})\|_2.$$

By lemma 4 and using $\mathsf{C}_{\mathsf{A}_{\mathsf{G}}} \cap \mathsf{G} = 0$, $|\mathsf{G}|^{-1/2}\boldsymbol{W}_{\mathsf{A}_{\mathsf{G}},\mathsf{G}}$ is a unitary transformation on the space $\boldsymbol{v}_{\mathsf{G}}$, so, using a standard norm equivalence,

$$\|\hat{\boldsymbol{p}}_{\mathsf{A}_{\mathsf{G}}} - \boldsymbol{p}_{\mathsf{A}_{\mathsf{G}}}\|_2 \leq |\mathsf{G}|^{-1/2}\|\hat{\boldsymbol{r}}_{\mathsf{G}} - \boldsymbol{r}_{\mathsf{G}}\|_2 \leq \|\hat{\boldsymbol{r}}_{\mathsf{G}} - \boldsymbol{r}_{\mathsf{G}}\|_\infty \leq \epsilon_0. \tag{37}$$

Under the conditions of proposition 8, $\epsilon_0 \leq O(\epsilon)\|\boldsymbol{r}_{\mathsf{G}}\|_\infty$. The bound on the number of rounds and the total number of measurements follows from the bound on the size of a stabilizer covering of G from lemma 2. □

In the special case that the group $\mathsf{G} = \mathsf{P}^n$, proposition 9 provides a direct method for estimating the Pauli projection of a noise channel and gives pointwise precision to within $O(\epsilon)(1 - p_0)$, where p_0 is the probability of no error. Because of its frequent use in applications, it may be desirable to state the following corollary of proposition 9 in terms of the diamond distance.

Corollary 10. *Let $\Lambda = \mathcal{E}^{\mathsf{P}^n}\big|_{\mathsf{G}}$ be the restriction to a subgroup $\mathsf{G} \subseteq \mathsf{P}^n$ of the Pauli projection of a channel $\mathcal{E}$. For any sufficiently small $\epsilon, \delta \geq 0$, an estimate $\hat{\Lambda}$ of any $\frac{1}{2}$-weak, $\frac{1}{2}$-stable, GTM noise model Λ can be reconstructed with*

$$\|\hat{\Lambda} - \Lambda\|_\diamond \leq O(\epsilon)\|\mathcal{I}_{\mathsf{G}} - \Lambda\|_F$$

using $O\Big[\frac{|\kappa||\mathsf{G}|^{1/2}}{\epsilon^2}\log\big(\frac{|\kappa||\mathsf{G}|}{\delta}\big)\Big]$ samples with $|\kappa| = O\big(\log(1/\Delta)\big)$ and sequence lengths at most $O(1/\Delta)$.

Proof. This follows from eq. (37) in proposition 9 using the norm equivalence

$$|\mathsf{G}|^{-1/2}\|\hat{\boldsymbol{p}}_{\mathsf{A}_\mathsf{G}} - \boldsymbol{p}_{\mathsf{A}_\mathsf{G}}\|_1 \leq \|\hat{\boldsymbol{p}}_{\mathsf{A}_\mathsf{G}} - \boldsymbol{p}_{\mathsf{A}_\mathsf{G}}\|_2 \,,$$

with $\|\hat{\boldsymbol{r}}_\mathsf{G} - \boldsymbol{r}_\mathsf{G}\|_2 \leq O(\epsilon)\|\boldsymbol{r}_\mathsf{G}\|_2$ from proposition 8 and $\|\boldsymbol{r}_\mathsf{G}\|_2 = \|\mathcal{I}_\mathsf{G} - \Lambda\|_F$, where $\mathcal{I}_\mathsf{G}$ is the restriction of the identity channel to G. □

VI. RECONSTRUCTING A SUBSET OF ERRORS

We have shown that all Pauli error rates for a marginal model can be reliably reconstructed with substantially fewer resources than might have been anticipated. However, reconstructing an exponential number of probabilities is manifestly inefficient. We now show how to reconstruct the dominant error rates in an approximately sparse noise model. The motivating application is that errors can be divided into a "background" noise process due to, e.g., T1/T2 lifetimes, and some additional errors that arise from unknown couplings. The goal is to learn these couplings and either determine the physical mechanism causing them and eliminate them via re-engineering the device or to introduce compensating pulses.

To set the scale for the number of dominant error rates, consider a background local depolarizing process on n qubits where independent single-qubit errors occur with probability p. Then a weight w error (that is, an error that acts nontrivially on w qubits) occurs with probability $\binom{n}{w}(1 - p)^{n-w}p^w$. For $np \ll 1$ (where, e.g., 50 qubits with $p = 0.001$ is approximately the current state of the art), the probability of an error with more than weight 1 is approximately $(np)^2/2$. Now suppose an additional noise process is mixed in that, with probability p' uniformly at random applies one of t additional errors with no *a priori* structure, representing, e.g., unknown correlated errors. Then $3n + t$ errors account for almost all of the errors in the device.

We first show how to use the **Ratio** subroutine to estimate the probabilities of all errors within an arbitrary set E. Formally, we will prove pointwise precision to within $O(\epsilon)(1 - p_0)$, where p_0 is the probability of no error. By eq. (21) we have $1 - p_0 = (1 + 1/d)r$ where r is the so-called average error rate. We will then present a divide-and-conquer routine to efficiently identify a set E of fixed size that can be used to identify correlated errors. The informal version of the following theorem stated as result 2 in the introduction follows from using the trivial stabilizer covering O of X with $|\mathsf{O}| = |\mathsf{X}|$.

Theorem 11. *For any set $\mathsf{E} \subseteq \mathsf{P}^n$, an estimator $\hat{\boldsymbol{p}}_\mathsf{E}$ of $\boldsymbol{p}_\mathsf{E}$ satisfying*

$$\|\hat{\boldsymbol{p}}_\mathsf{E} - \boldsymbol{p}_\mathsf{E}\|_\infty \leq O(\epsilon)(1 - p_0)$$

with probability at least $1 - \delta$ can be obtained from the output of **Ratio**$(\mathsf{O}, \mathsf{X}, t)$ *using a random set $\mathsf{X} \subseteq \mathsf{P}^n$ with $|\mathsf{X}| = \frac{1}{\epsilon^2}\log(4|\mathsf{E}|/\delta)$, any stabilizer covering O of X, and $t = \frac{1}{\epsilon^2}\log(4|\mathsf{X}||\kappa|/\delta)$ measurements per round for $|\mathsf{O}|$ rounds.*

Proof. By eq. (18) and lemmas 1 and 4, we have

$$\begin{aligned} p_a = (a|\boldsymbol{W}^{-1}\boldsymbol{f} &= \frac{1}{|\mathsf{P}^n|}\boldsymbol{W}_{a,\mathsf{P}}\boldsymbol{f} \\ &= \mathbb{E}_{b\in\mathsf{P}^n}\left[(-1)^{\langle a,b\rangle}f_b\right] \\ &= 1[a=0] - \mathbb{E}_{b\in\mathsf{P}^n}\left[(-1)^{\langle a,b\rangle}r_b\right] \end{aligned}$$

for any $a \in \mathsf{P}^n$. Let $\mathsf{X} \subseteq \mathsf{P}^n$ be a set with $|\mathsf{X}|$ elements sampled independently and uniformly at random without replacement and

$$\tilde{p}_a = 1[a=0] - \mathbb{E}_{b\in\mathsf{X}}\left[(-1)^{\langle a,b\rangle}r_b\right].$$

By [21, Lemma 4], we have $r_b \in [0, 2(1-p_0)]$ for all $b \in \mathsf{P}^n$. Applying Serfling's inequality [67] to the first term gives

$$\Pr\big(|\tilde{p}_a - p_a| \geq \epsilon(1-p_0)\big) \leq 2\mathrm{e}^{-|\mathsf{X}|\epsilon^2/2(1-|\mathsf{X}|/|\mathsf{P}^n|)} \leq 2\mathrm{e}^{-|\mathsf{X}|\epsilon^2/2} =: \delta/2. \tag{38}$$

Now suppose we have estimates $\hat{r}_b$ of r_b for all $b \in \mathsf{X}$ and let

$$\hat{p}_a = 1[a=0] - \mathbb{E}_{b\in\mathsf{X}}\left[(-1)^{\langle a,b\rangle}\hat{r}_b\right]. \tag{39}$$

By the triangle inequality and eq. (38), we have that with probability at least $1-\delta/2$,

$$|\hat{p}_a - p_a| \leq \epsilon(1-p_0) + \left|\mathbb{E}_{b\in\mathsf{X}}(-1)^{\langle a,b\rangle}(\hat{r}_b - r_b)\right|.$$

By proposition 8, $|\hat{r}_b - r_b| \leq O(\epsilon)r_b$ with probability at least $1-\delta/2$ using $t = \frac{2}{\epsilon^2}\log(\frac{4|\mathsf{X}||\kappa|}{\delta})$ measurements per round. Therefore by the union bound, we have that with probability at least $1-\delta$,

$$|\hat{p}_a - p_a| \leq O(\epsilon)(1-p_0)$$

for any $a \in \mathsf{P}^n$. The final result holds for any set $\mathsf{E} \subseteq \mathsf{P}^n$ by the union bound, redefining $\delta \to \delta/|\mathsf{E}|$. □

We now provide a search heuristic to identify sets E of interest using a function **Select** that returns the indices (as Pauli operators) of entries in a vector $\boldsymbol{p}$ of probabilities satisfying a desired condition. We use three additional subroutines in what follows. **Choose**(A, b) returns a set of $\min\{b, |\mathsf{A}|\}$ elements of A chosen uniformly at random without replacement. **Cover**(F) returns any valid stabilizer covering of a set $\mathsf{F} \subseteq \mathsf{P}^n$. **Ratio** is as described in section IV. For any $\mathsf{A} \subset \mathbb{N}$, let $\mathsf{P}_{\mathsf{A}} \cong \mathsf{P}^{|\mathsf{A}|}$ be the set of Pauli operators that act trivially on all qubits not in A and the *support* of a set $\mathrm{supp}(\mathsf{X}) \subseteq \mathsf{P}^n$ be the set of qubits that are acted on nontrivially by some element of X.

```
TreeReconstruction($t$, Select, $u$, $n$) Reconstruct an $n$-qubit error model using a
function Select to select error terms at each stage of the iteration.
  1. Set $E' := \{\mathsf{P}_{\{j\}} : j \in \mathbb{Z}_n\}$.
  2. For $j = 0, 1, \ldots, \lceil \log_2 n \rceil$, Do
     (a) Set $E := E'$.
     (b) Set $F := \emptyset$.
     (c) For each $e \in E$, Do
           • Set $F := F \cup$ Choose$(\mathsf{P}_{\mathrm{supp}(e)}, u)$.
     (d) Set $\mathsf{O} :=$ Cover$\big(\cup_{f\in F} f\big)$.
     (e) Set $\boldsymbol{r} :=$ Ratio$(\mathsf{O}, \cup_{f\in F} f, t)$.
     (f) Set $E' := \emptyset$.
```

```
    (g) For $k = 0, 1, \ldots, |E| - 1$, Do
        • Set $\boldsymbol{p} := \sum_{e \in E_k} \mathbb{E}_{x \in F_k} (-1)^{\langle e,x \rangle} r_x |e)$.
        • Set $E_k :=$ Select$(\boldsymbol{p})$.
        • If $k$ is odd, Set $E' := E' \cup \{E_{k-1} \otimes E_k\}$.
    (h) If $|E|$ is odd, set $E' := E' \cup \{E_{|E|-1}\}$.
3. Return $\boldsymbol{p} := \sum_{e \in E_0} \mathbb{E}_{x \in F_0} (-1)^{\langle e,x \rangle} r_x |e)$.
```

Theorem 12. *Let* **Select** *be any function, u be any positive integer and* **Cover**(X) *return the trivial cover from lemma 2. Then for sufficiently large t,* **TreeReconstruction**$(t, \mathbf{Select}, u, n)$ *uses $O\big(utn \log(1/\Delta)\big)$ measurements.*

Proof. At step 2(c) in the jth iteration, we are guaranteed that each of the at most 2^j subsets of F each contain at most u terms. Therefore $| \cup_{f \in \mathsf{F}} f| = |\mathsf{O}| \leq 2^{\lceil \log_2(n) \rceil - j} u$ in each iteration. For sufficiently large t, **Ratio**$(\mathsf{O}, \cup_{f \in F} f, t)$ uses t measurements for each of $|\mathsf{O}||\kappa|$ rounds where $|\kappa| = O\big(\log \frac{1}{\Delta}\big)$. Therefore the jth iteration uses $O\big(2^{\lceil \log_2(n) \rceil - j} ut \log \frac{1}{\Delta}\big)$ measurements. Summing over $j = 0, \ldots, \lceil \log_2 n \rceil$ completes the proof. □

A natural choice of **Select** is the function that returns the indices of the s largest elements of the input $\boldsymbol{p}$. With this function, **TreeReconstruction** will heuristically return a set E that is a good s-sparse approximation of the Pauli channel. We note that we do not attempt to prove that the resulting set is a best s-sparse approximation, that is, that it is a maximizer (or near-maximizer) of $\|\boldsymbol{p}_{\mathsf{E}}\|_1$ over all E of size s. It would be possible to prove an upper bound on how close it gets to an optimal approximation by estimating the precision of each error in the set to high precision and backtracking if $\sum_{e \in E_k} \boldsymbol{p}_e$ is below some threshold value at any iteration. However, a rigorous statement along these lines would likely only hold with impractical amounts of resources or a tighter analysis of the distribution of the output of **Ratio**. We leave the question of provable approximation ratios to future work.

We also note that a better run-time may be achieved in this setting using methods for reconstructing a sparse signal from few measurements [68–70]. Specifically, the present variant can be cast in a similar form to sparse Fourier and Hadamard transforms [71–74]. However, it is an open question if the methods in those references can be adapted to the symplectic structure of the Pauli group to achieve near-optimal sparse reconstruction. Moreover, while finding such a set is desirable, we do not require any property of the set E for the reconstruction of $\boldsymbol{p}_{\mathsf{E}}$ to be pointwise convergent. Thus, when the heuristic converges to a sparse set with large measure it is a certificate of correctness since each estimate in the set is pointwise accurate.

Another interesting choice of **Select** is a function that returns the indices of errors that do not fit a background model. The most natural background model is for independent single-qubit errors, which can be directly estimated using theorem 11.

VII. BOUNDED DEGREE GRAPHICAL MODELS

In this section we show how the reconstruction procedures above for complete Pauli channels can be used to learn a Pauli channel on n qubits with bounded degree correlations. It will be convenient to adopt a slightly different notation than what was used previously that treats the probability distribution p over Pauli errors as a function of a random variable rather than a vector. That is, we write $p(\boldsymbol{x})$ for the probability of the string of Paulis $\boldsymbol{x}$, $p(\boldsymbol{x}|\boldsymbol{y})$ for the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{y}$, and so on. In particular, if $p(\boldsymbol{x}, \boldsymbol{y})$ is a joint distribution, then we write $p(\boldsymbol{x})$ for the marginal. We will now review some concepts from the field of probabilistic graphical models; see Ref. [75] for an introduction.

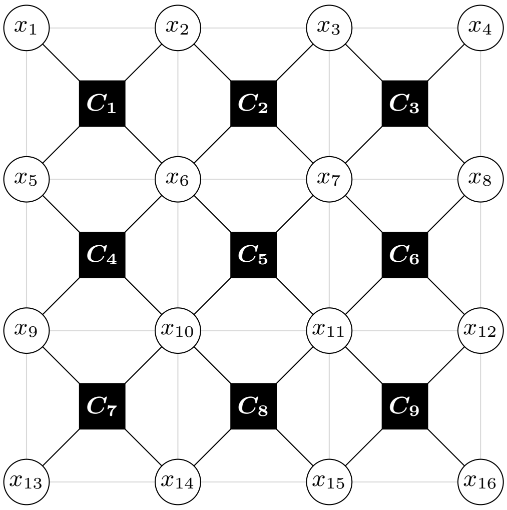

FIG. 1. An example of a factor graph. The factors are the black squares and the variable nodes are the white circles. This is also an example of a bounded degree factor graph because each each variable belongs to at most 4 factors, and each factor couples only 4 variables.

Probabilistic graphical models are, at the most basic level, probability distributions over collections of random variables where the dependency structure between the variables is specified by a bipartite graph called a factor graph. Consider a collection of random variables with x_j denoting the jth random variable. The variable $\boldsymbol{x}$ will be a random n-qubit Pauli, with x_j being the jth tensor factor (i.e., the single-qubit Pauli acting on qubit j), but we leave the discussion more general for the moment. Suppose that there exist *factors* C_k that are subsets of the random variables such that the joint probability distribution p over all the x_j obeys

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_k \phi_k(\boldsymbol{x}_{C_k}), \tag{40}$$

where $\phi_k : \boldsymbol{x}_{C_k} \to \mathbb{R}^+$ are strictly positive functions called *factor potentials* supported on the factors; the argument of ϕ_k is shorthand notation for the subset of variables contained in a factor, $\boldsymbol{x}_C = \{x_j : j \in C\}$; and Z is a normalization constant called the *partition function.* (Note that some authors use the term factor potential to refer instead to $\log \phi_k$.) Then the factor graph for this model is the bipartite graph with one set of nodes labeled by the variable labels j and the other set of nodes labeled by the factors C_k, and an edge between j and C_k if and only if $x_j \in C_k$. Such a factorized strictly positive probability distribution is known as a *Gibbs random field* since the factor potentials play the role of exponentiated energy potentials in the Gibbs distribution of a statistical mechanical model. Note that there always exists a Gibbs random field for any probability distribution, although the factor graph may be trivial, that is, contain only one factor.

The structure of a nontrivial factor graph implies that not all variables in $\boldsymbol{x}$ can be arbitrarily correlated. In particular, the correlation is controlled by the *Markov blanket* $\boldsymbol{x}_{\partial S}$ of a subset of variables $\boldsymbol{x}_S$, where the set ∂S is defined as

$$\partial S = \bigcup_k \{C_k : C_k \cap S \neq \emptyset\} - S. \tag{41}$$

That is, it is the set of variables that are adjacent to S via a factor node in the factor graph, but excluding S itself. We will refer to both ∂S and $\boldsymbol{x}_{\partial S}$ as the Markov blankets of both S and $\boldsymbol{x}_S$ with context resolving any ambiguity. The closure $\bar{S}$ is the union of a set and its Markov blanket, $\bar{S} = S \cup \partial S$.

The celebrated Hammersley-Clifford theorem [38, 76] describes how the Markov blanket controls the dependency between variables in a Gibbs random field. The variables $\boldsymbol{x}_A$ and $\boldsymbol{x}_{A^c}$, where A^c is the complement of A, obey the following conditional independence relation in terms of the Markov blanket of A,

$$p(\boldsymbol{x}_A|\boldsymbol{x}_{A^c}) = p(\boldsymbol{x}_A|\boldsymbol{x}_{\partial A}). \tag{42}$$

This relation, known as the *local Markov property*, states that the variables in a set A have only bounded correlation in the sense that they are conditionally independent of any other variables outside of their Markov blanket ∂A.

To illustrate some of these concepts, consider the example factor graph given in fig. 1. Here the factors are the neighborhoods of the black squares, so for example $C_1 = \{x_1, x_2, x_5, x_6\}$. The Markov blanket of x_1 is given by $\{x_2, x_5, x_6\}$, and the Markov blanket of $\{x_1, x_5\}$ is $\{x_2, x_6, x_9, x_{10}\}$. The closure of $\{x_1, x_2\}$ is $\{x_1, x_2, x_3, x_5, x_6, x_7\}$. The variable x_1 is conditionally independent of every variable except $\{x_2, x_5, x_6\}$ and the set of random variables $\{x_6, x_7, x_{10}, x_{11}\}$ depends conditionally on every other variable.

We will provide an estimate of a Gibbs random field by estimating the factor potentials. We will not provide estimates for the partition function because estimating general partition functions is believed to be computationally hard [77, 78], and so properly normalizing our probability estimate will in general be intractable. However, there are still several cases of practical interest where this is provably not the case. For example, when the factor graph is a tree, and when $p(\mathbf{0})$ is not too small to be estimated via sampling. Even without being able to efficiently compute the normalization exactly, either heuristics could be used, or ratios of probabilities can be used unconditionally. Ratios of probabilities are all that is needed in, for example, most Monte Carlo methods. However, as we now show, the error on the renormalized distribution can be bounded by the error on the factor potentials without needing to obtain the partition functions.

Lemma 13. *Let p and q be Gibbs random fields with the same factor graph and factorizations $p(\boldsymbol{x}) = \frac{1}{Z_p}\prod_C \phi_{p,C}(\boldsymbol{x}_C)$ and $q(\boldsymbol{x}) = \frac{1}{Z_q}\prod_C \phi_{q,C}(\boldsymbol{x}_C)$ respectively. Then*

$$\|p-q\|_1 \leq \sum_C \max_{\boldsymbol{x}_C} |\log \tfrac{\phi_{p,C}(\boldsymbol{x}_C)}{\phi_{q,C}(\boldsymbol{x}_C)}|.$$

Proof. We begin by recalling Pinsker's inequality [79], one form of which is given by

$$\|p-q\|_1^2 \leq D(p\|q) + D(q\|p) \tag{43}$$

where $D(p\|q)$ is the relative entropy, defined by

$$D(p\|q) = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \tfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}.$$

Thus it suffices to bound the symmetric relative entropy, which simplifies to

$$D(p\|q) + D(q\|p) = \sum_{\boldsymbol{x}} [p(\boldsymbol{x}) - q(\boldsymbol{x})] \log \tfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$$

for any probability distributions p and q. Substituting the definitions of the Gibbs random fields into the logarithms, we have

$$\begin{aligned} D(p\|q) + D(q\|p) &= \sum_{\boldsymbol{x}} [p(\boldsymbol{x}) - q(\boldsymbol{x})] \left[\log \tfrac{Z_p}{Z_q} + \sum_C \log \tfrac{\phi_{p,C}}{\phi_{q,C}}\right] \\ &= \sum_C \sum_{\boldsymbol{x}} [p(\boldsymbol{x}) - q(\boldsymbol{x})] \log \tfrac{\phi_{p,C}}{\phi_{q,C}}, \end{aligned}$$

where to obtain the second line we have used the fact that Z_p and Z_q are independent of $\boldsymbol{x}$ and that $\sum_{\boldsymbol{x}} p(x) - q(x) = 0$, which holds as p and q are probability distributions. By applying Holder's inequality separately for each C, we obtain

$$\begin{aligned} D(p\|q) + D(q\|p) &\leq \sum_C \|p-q\|_1 \max_{\boldsymbol{x}} |\log \tfrac{\phi_{p,C}}{\phi_{q,C}}| \\ &\leq \sum_C \|p-q\|_1 \max_{\boldsymbol{x}_C} |\log \tfrac{\phi_{p,C}}{\phi_{q,C}}|, \end{aligned}$$

where the second line follows from the bounded dependence of the factors. Substituting the bound on the symmetric relative entropy into eq. (43) and canceling the common factor of $\|p-q\|_1$ gives the desired result. □

Given a probability distribution $p(\boldsymbol{x})$ that is a Gibbs random field, the choice of factor potentials and partition function are not unique. However, the Hammersley-Clifford theorem also gives an explicit description of a certain canonical choice for the factor potentials as a function of the marginal probability distributions on the factors and their Markov blankets, as well as an explicit partition function. It is this description that we will use for our estimators below. We use the parameterization from Abbeel *et al.* [57] that applies directly to factor graphs (as opposed to less general formulations like Bayesian networks or Markov random fields) and makes explicit use of the local Markov property. We first fix a fiducial assignment to $\boldsymbol{x}$ which we call $\mathbf{0}$. This is simply a reference value and can be any arbitrary fixed assignment to $\boldsymbol{x}$; however we will see below that a convenient choice for our purposes will be to choose as a reference value the "identity Pauli" outcome for each variable. Next we will augment the factors in our graph and consider 2^{C_k}, the set of all subsets of C_k. The union of all of these, minus the empty set, defines a new, larger collection of factors C^*, given explicitly by

$$\mathsf{C}^* = \bigcup_k 2^{C_k} - \emptyset. \tag{44}$$

Next, given a factor $C_k^* \in \mathsf{C}^*$, define the *canonical factor potential* for C_k^* by

$$\log \phi_k(\boldsymbol{x}_{C_k^*}) = \sum_{S \subseteq C_k^*} (-1)^{|C_k^*|-|S|} \log p(\boldsymbol{x}_S|\mathbf{0}_{\partial S}). \tag{45}$$

Let C_j be a factor such that $C_k^* \in 2^{C_j}$. Then every $S \subset C_k^*$ is also a subset of C_j and so every probability on the right-hand-side of eq. (45) depends only upon $\bar{C}_j$, the closure of C_j with its Markov blanket. That is, for all S in the sum we have

$$S \subseteq C_k^* \cup \partial C_k^* \subseteq C_j \cup \partial C_j = \bar{C}_j. \tag{46}$$

Therefore each ϕ_k is determined by $p(\boldsymbol{x}_{\bar{C}_j})$ for some factor C_j. If each $\bar{C}_j$ has constant size, each $\log \phi_k$ is completely determined by a small amount of data.

The Gibbs random field can now be expressed in terms of the canonical factor potentials as

$$p(\boldsymbol{x}) = p(\mathbf{0}) \prod_{C_j^* \in \mathsf{C}^*} \phi_j(\boldsymbol{x}_{C_j^*}). \tag{47}$$

We note that $1/p(\mathbf{0})$ plays the role of the partition function Z and can be independently estimated via theorem 11.

The error bounds below are derived assuming that an independent estimate is obtained for each of the raw factors in the canonical factor decomposition. This adds a constant factor overhead to the overall sample complexity (when the degree of the factor graph is bounded). This overhead could be avoided by using estimates from a covering set of marginals and then directly computing estimates of the marginals on the subfactors, that is, on the subsets of the factors. However, this might bias the overall estimate somewhat and would complicate the analysis presented below, so we leave an understanding of this more efficient protocol to future work.

Importantly, the estimates for the individual raw factors will not generally be self-consistent, and expressing this requires careful notation. Let $\hat{p}_S(\boldsymbol{x}_S)$ denote an empirical estimate of the marginal distribution $p(\boldsymbol{x}_S)$ for a set S, which in our case will be obtained from proposition 9. Then for two sets A, B such that $S \subset A$, $S \subset B$ and $A \neq B$, the two marginal distributions obtained over S from the empirical estimates do not need to agree, that is,

$$\sum_{\boldsymbol{x}_{A-S}} \hat{p}_A(\boldsymbol{x}_{A-S}, \boldsymbol{x}_S) \neq \sum_{\boldsymbol{x}_{B-S}} \hat{p}_B(\boldsymbol{x}_{B-S}, \boldsymbol{x}_S)\,. \tag{48}$$

Indeed, they will generically not agree in the presence of sampling errors. Moreover, neither will generally agree with the corresponding marginal of the global reconstruction q obtained by substituting the $\hat{p}_S$ into eq. (45). Thus, the canonical factor potentials are essential to round each of the local empirical marginals into a global coherent probability distribution with self-consistent local marginals.

As a final difficulty, we cannot use an estimate of $p(\mathbf{0})$ as the partition function to exactly normalize the empirically reconstructed distribution. Fortunately, as shown in lemma 13, we do not need to know the exact value of the partition function for the empirical distribution to bound the error between our empirical estimate and the true distribution.

Our estimator of the global Gibbs distribution is as follows. We will simply use the expressions in eqs. (45) and (47) where empirical estimates of the local conditional probabilities are used in place of the exact distributions. These can be obtained by using proposition 9 to estimate the complete marginal on the closure of each factor, $\bar{C}$. When an estimate q is obtained in this way, we say that q is a *canonical estimator* of p.

Definition 4 (Canonical estimator)**.** Given a Gibbs random field p with known factor graph C, a distribution $q = q\big(\{\hat{p}_{\bar{C}} : C \in \mathsf{C}^\star\}\big)$ obtained by substituting empirical estimates $\{\hat{p}_{\bar{C}} : C \in \mathsf{C}^\star\}$ into eqs. (45) and (47) is called a *canonical estimator* of p.

We now bound the error on our empirical estimate assuming it has been properly normalized. In order to obtain a nontrivial bound, we require that the empirical estimates of the marginal probabilities used to estimate the factor potentials are sufficiently close to their true values. We also require that the true and estimated values are strictly positive, a point that we will quantify below. Finally, we will also use the bound on the number of variables that are in any given factor, defining

$$\nu \coloneqq \max_{C\in\mathsf{C}} |C| \tag{49}$$

to be the maximum degree of all the factors.

Lemma 14. *Let p be a Gibbs random field with a factor graph C and factorization $\frac{1}{Z_p}\prod_{C\in\mathsf{C}}\phi_{p,C}$, let N be the number of factors in C, and let ν be the maximum degree of all the factors. Then any canonical estimator q satisfies*

$$\|p-q\|_1 \le N3^\nu \max_{\boldsymbol{x}_C}\Big|\log \tfrac{p(\boldsymbol{x}_C|\mathbf{0}_{\partial C})}{\hat{p}_{\bar{C}}(\boldsymbol{x}_C|\mathbf{0}_{\partial C})}\Big|, \tag{50}$$

where the maximum is over all variables $\boldsymbol{x}_C$ in either a factor or a subfactor in C.

Proof. Applying lemma 13 to the canonical factor decomposition gives

$$\|p-q\|_1 \le \sum_{C_k\in\mathsf{C}}\sum_{S\subseteq C_k}\max_{\boldsymbol{x}_S}\Big|\log\tfrac{\phi_{p,S}(\boldsymbol{x}_S)}{\phi_{q,S}(\boldsymbol{x}_S)}\Big|.$$

Using the canonical factor potentials from eq. (45), we have

$$(-1)^{|S|}\log\tfrac{\phi_{p,S}(\boldsymbol{x}_S)}{\phi_{q,S}(\boldsymbol{x}_S)} = \sum_{R\subseteq S}(-1)^{|R|}\log\tfrac{p(\boldsymbol{x}_R|\mathbf{0}_{\partial R})}{\hat{p}_{\bar{R}}(\boldsymbol{x}_R|\mathbf{0}_{\partial R})}\,.$$

Taking the absolute value and using the triangle inequality over the $2^{|S|}$ subsets of S, we obtain

$$\begin{aligned}\|p-q\|_1 &\le \sum_{C_k\in\mathsf{C}}\sum_{S\subseteq C_k}\sum_{R\subseteq S}\max_{\boldsymbol{x}_R}\Big|\log\tfrac{p(\boldsymbol{x}_R|\mathbf{0}_{\partial R})}{\hat{p}_{\bar{R}}(\boldsymbol{x}_R|\mathbf{0}_{\partial R})}\Big| \\ &\le N3^\nu\max_{\boldsymbol{x}_C}\Big|\log\tfrac{p(\boldsymbol{x}_C|\mathbf{0}_{\partial C})}{\hat{p}_{\bar{C}}(\boldsymbol{x}_C|\mathbf{0}_{\partial C})}\Big|\,,\end{aligned}$$

where in the last line we used the bound on the number of factors N, the bound on the degree of the factors $|C_k| \le \nu$, and the identity $\sum_{A\subseteq B\subseteq C} 1 = 3^{|C|}$. The maximization in the last line is over all variables $\boldsymbol{x}_C$ where C is a factor or a subfactor, and the result is immediate. □

The next lemmas let us translate the previous bound in terms of the logarithms of the conditional marginals and their empirical estimates into a bound in terms of the quantities that we naturally have control over, namely the (non-conditional) marginals and their empirical estimates.

Lemma 15. *For $ab > 0$, $\left|\log \frac{a}{b}\right| \le \frac{|a-b|}{\sqrt{ab}}$.*

Proof. This result is a consequence of a simpler inequality in terms of a single variable. For $x \ge 1$ we can use the integral representation of $\log x$ and the Cauchy-Schwarz inequality to obtain

$$\log x = \int_1^x \frac{1}{z} \mathrm{d}z \le \left(\int_1^x \mathrm{d}z\right)^{1/2} \left(\int_1^x \frac{1}{z^2} \mathrm{d}z\right)^{1/2} = \sqrt{x} - \frac{1}{\sqrt{x}} \qquad (x \ge 1).$$

Multiplying by -1, we find

$$-\log x = \log \tfrac{1}{x} \ge -\sqrt{x} + \frac{1}{\sqrt{x}},$$

and by letting $0 < z = 1/x \le 1$, we see that the reverse inequality is true in the interval $x \in (0, 1]$. This establishes that for all $x > 0$,

$$|\log x| \le \left|\sqrt{x} - \tfrac{1}{\sqrt{x}}\right|.$$

The lemma follows by letting $x = a/b$ for $ab > 0$ and using some basic algebra. □

Lemma 16. *For any two strictly positive probability distributions $p(x, y)$, $q(x, y)$ on alphabets X, Y and for any fixed elements $x \in X$ and $y \in Y$ we have*

$$\left|\log \tfrac{p(x|y)}{q(x|y)}\right| \le \frac{|p(x, y) - q(x, y)| + |p(y) - q(y)|}{p(y)\, \sqrt{p(x|y)q(x|y)}}.$$

Proof. By lemma 15, we have

$$\left|\log \tfrac{p(x|y)}{q(x|y)}\right| \le \frac{|p(x|y) - q(x|y)|}{\sqrt{p(x|y)q(x|y)}}.$$

From the definition of conditional probability, the triangle inequality, and the fact that p and q are probability distributions, we then have

$$\begin{aligned}|p(x|y) - q(x|y)| &= \frac{1}{p(y)} \big|p(x, y) - q(x, y) + q(x|y)[q(y) - p(y)]\big| \\ &\le \frac{|p(x, y) - q(x, y)| + |q(y) - p(y)|}{p(y)}.\end{aligned}$$

Combining these two inequalities yields the desired bound. We remark that this proof works equally well with $p \leftrightarrow q$, so the $p(y)$ in the denominator could also be replaced with $\max\{p(y), q(y)\}$ to get a stronger bound. □

The following definitions help facilitate a direct application of proposition 8 to the problem of learning the parameters of a Pauli channel with a bounded-degree factor graph. Here, as before, the minimizations in these definitions are over every factor or subfactor in the factor graph of p and q. We first introduce a quantitative notion of positivity given by the minimum geometric mean marginal probability over a factor C, defined by

$$G(p, q) := \min_{\boldsymbol{x}_C} \sqrt{p(\boldsymbol{x}_C|\mathbf{0}_{\partial C})q(\boldsymbol{x}_C|\mathbf{0}_{\partial C})} \tag{51}$$

for the fixed but arbitrary outcome $\mathbf{0}_{\partial C}$. We will also quantify the deviation of the local marginals from their empirical estimates by introducing the following error parameters. Let

$$\epsilon_1 := \max_{\boldsymbol{x}_C}\left|p(\boldsymbol{x}_C, \mathbf{0}_{\partial C}) - \hat{p}_{\bar{C}}(\boldsymbol{x}_C, \mathbf{0}_{\partial C})\right| \quad \text{and} \quad \epsilon_2 := \max_{\partial C}\left|p(\mathbf{0}_{\partial C}) - \hat{p}_{\bar{C}}(\mathbf{0}_{\partial C})\right|. \tag{52}$$

For Pauli channels with maximum degree ν, the triangle inequality shows that $\epsilon_2 \le 4^\nu \epsilon_1$. Finally, we introduce

$$\gamma := \min_{\partial C} p(\mathbf{0}_{\partial C})\,. \tag{53}$$

From these definitions, we have the following lemma.

Lemma 17. *Let p be a Gibbs random field with a factor graph C and factorization $\frac{1}{Z_p}\prod_{C\in\mathsf{C}}\phi_{p,C}$, let N be the number of factors in C, and let ν be the maximum degree of all the factors. Then any canonical estimator q satisfies*

$$\|p-q\|_1 \le \frac{N3^\nu}{\gamma G}(\epsilon_1+\epsilon_2), \tag{54}$$

where G is as in eq. (51), ϵ_1, ϵ_2 are as in eq. (52), and γ is as in eq. (53).

Proof. The result follows immediately by combining lemma 14 with lemma 16 and the definitions in eqs. (51) to (53). □

Lemma 17 can be used together with our procedure for learning local marginal distributions (that is, complete Pauli noise models on a subset of qubits) to obtain a global guarantee on the 1-norm error between our estimate and the true distribution. As we see in the lemma, the quality of the estimate will depend on two quantities related to the local marginals as well as the precision of our local estimate, ϵ from proposition 9. The first additional quantity is the local marginal that is furthest from the noiseless case, as quantified by the 2-norm,

$$r^\star = \max_C \|\mathbf{1}_I - p(\boldsymbol{x}_{\bar{C}})\|_\infty. \tag{55}$$

Note that by eq. (21) we have $\|\mathbf{1}_I - p(\boldsymbol{x}_{\bar{C}})\|_\infty \le \|\mathbf{1}_I - p(\boldsymbol{x}_{\bar{C}})\|_1 = 2\frac{d+1}{d}r \le 3r$ where r is the average error rate of the local noise channel supported on $\bar{C}$, so we expect that this term is small. The second is the minimum geometric mean marginal probability from eq. (51). Finally, the geometry of the factor graph is also important. We will need to use the bounded degree assumption to get a bound on $|\bar{C}|$ that is independent of n and N. We will quantify the geometry of the factor graph by defining, in analogy with the definition of ν from eq. (49),

$$\bar{\nu} := \max_C |\bar{C}|\,. \tag{56}$$

In general $\bar{\nu}$ can depend on both n and N, but it is $O(1)$ for factor graphs having factors of bounded size and where each variable participates in a bounded number of factors. From these quantities, we have the following global guarantee.

Proposition 18. *Let p be a Gibbs distribution with a known factor graph C having N factors, n variables, and $\bar{\nu} = O(1)$ such that p corresponds to the Pauli error rates for a quantum channel. Suppose that this noisy channel is GTM, $\frac{1}{2}$-weak, $\frac{1}{2}$-stable, and that $\Delta_{\bar{C}} = O(1)$. Then there exists a canonical estimator q such that for all sufficiently small $\epsilon > 0$, using T samples with*

$$T = O\left(\frac{N^2}{G^2\epsilon^2}\log\left(\frac{N}{\delta}\right)\right) \tag{57}$$

we have, with probability at least $1-\delta$,

$$\|p-q\|_1 \le \epsilon r^\star\,. \tag{58}$$

Moreover, an estimate proportional to q can be found in time $\mathrm{poly}(n)$.

Proof. We construct the following canonical estimator q. Using **RunCB** and **Ratio**, we obtain estimates of the marginal distribution $\hat{p}_{\bar{C}}(\boldsymbol{x}_{\bar{C}})$ for every factor C of the graph. By proposition 9, each of these estimates can be obtained such that

$$\|\hat{p}_{\bar{C}}(\boldsymbol{x}_{\bar{C}}) - p_{\bar{C}}(\boldsymbol{x}_{\bar{C}})\|_2 \le O(1)\epsilon r^\star\,,$$

with probability δ using $T_{\bar{C}} = O\big(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\big)$ samples per marginal. Here the big-O notation hides factors of $\kappa = O\big(\log\frac{1}{\Delta_{\bar{C}}}\big)$ where $\Delta_{\bar{C}}$ is the spectral gap of the marginal channel as defined in eq. (24), but these are $O(1)$ by assumption. By the triangle inequality, these bounds apply to any marginals that are computed from these estimates as well, so they can also be applied to the subfactors in the canonical factor decomposition.

The above bounds for the empirical marginals imply that ϵ_1 and ϵ_2 from eq. (52) are also both bounded as $\epsilon_1, \epsilon_2 \leq O(1)\epsilon r^\star$ with probability at least $1 - N\delta$. Because the noise is assumed to be $1/2$-weak, we also have that γ from eq. (53) is bounded as $\gamma \geq 1/2$. Now applying lemma 17, we have a canonical estimator q such that with probability at least $1-\delta$

$$\|p - q\|_1 \leq O(1)\frac{N}{G}\epsilon r^\star \tag{59}$$

using $T = O\big(\frac{N}{\epsilon^2}\log\frac{N}{\delta}\big)$ samples.

This sample complexity can be improved by a factor of N as follows. Consider the graph whose vertices are the closures of each factor, $\bar{C}$, and whose edges are present if and only if $\bar{C}_1 \cap \bar{C}_2 \neq \emptyset$. Because the factor graph is assumed to have constant degree, this adjacency graph also has constant degree, and so the $\bar{C}$ vertices can be partitioned into a constant number of independent sets. By construction, nonadjacent factors are conditionally independent, and so they can be estimated from the same sample. That is to say, each independent set can be chosen to comprise $\Omega(N)$ vertices, and the constituent marginals associated to these factors can be estimated in parallel. Such a choice can be found in polynomial time in N using, e.g., a greedy algorithm. The estimation is repeated across the $O(1)$ distinct independent sets, which reduces the sample complexity by a factor of $O(N)$, as claimed.

The final result is obtained by rescaling ϵ to ϵ/N. As all the factors are assumed to have size $O(1)$, the canonical estimator q can be constructed in polynomial time in n (or equivalently N) except for the normalizing factor. □

Finally, we demonstrate that significantly better bounds that scale as $\sqrt{n}$ can be obtained in terms of the differences between the marginal distributions of p and q. The following bound is difficult to apply theoretically because we cannot directly compute or bound the marginals of the empirical distribution without first computing or bounding the partition function for the reconstructed distribution. However, the marginals of the true probability distribution can be accurately estimated by proposition 8 and the marginals of the empirical distribution can be estimated heuristically by Monte Carlo sampling from the factor potentials.

Lemma 19. *Given two Gibbs random fields p and q on n variables with the same factor graph, we have the bound*

$$\|p - q\|_1 \leq \sqrt{n}\max_u \frac{1}{\sqrt{q_u^\star}}\big\|p(\boldsymbol{x}_{\bar{u}}) - q(\boldsymbol{x}_{\bar{u}})\big\|_2, \tag{60}$$

where $q_u^\star = \min_{\boldsymbol{x}_{\bar{u}}} q(\boldsymbol{x}_{\bar{u}})$ and $\bar{u}$ denotes the closure $u \cup \partial u$ of u.

Proof. We will bound the relative entropy and then use Pinsker's inequality. Recall the chain rule for relative entropy [80],

$$D\big(p(x,y|z)\|q(x,y|z)\big) = D\big(p(y|z)\|q(y|z)\big) + D\big(p(x|y,z)\|q(x|y,z)\big).$$

By iteratively applying the chain rule to each variable, we have

$$\begin{aligned} D\big(p\|q\big) &= \sum_{u=1}^{n} D\big(p(x_u|\boldsymbol{x}_{>u})\|q(x_u|\boldsymbol{x}_{>u})\big) \\ &= \sum_{u=1}^{n} D\big(p(x_u|\boldsymbol{x}_{\partial^+ u})\|q(x_u|\boldsymbol{x}_{\partial^+ u})\big), \end{aligned}$$

where in the second step we use the local Markov property to restrict the conditioning to the remaining part of the Markov blanket of u, which we denote $\boldsymbol{x}_{\partial^+ u}$. Because the relative entropy is positive, the chain rule also proves that marginalizing or conditioning on random variables only decreases the relative entropy. Therefore, we can add back the variables from the Markov blanket of u to get the upper bound

$$D\big(p\|q\big) \leq \sum_{u=1}^{n} D\big(p(\boldsymbol{x}_{\bar{u}})\|q(\boldsymbol{x}_{\bar{u}})\big).$$

Now we note that from $\log x \leq x - 1$ we have

$$D\big(p\|q\big) + D\big(q\|p\big) = \sum_x \Big(p(x) - q(x)\Big) \log \tfrac{p(x)}{q(x)} \leq \sum_x \frac{\big(p(x) - q(x)\big)^2}{q(x)} \leq \frac{1}{\min_x q(x)} \big\|p - q\big\|_2^2.$$

The result then follows by applying this bound term by term, taking a maximum, and then substituting into eq. (43). □

VIII. CONCLUSION

We have shown that Pauli channels can be learned with high precision using far fewer resources than previous methods. As noise can be engineered to be effectively Pauli noise using randomized compiling [2], the methods described here should be broadly applicable to learning the residual noise under randomized compiling for large-scale devices, and, as such, represent a paradigm shift for characterizing quantum hardware.

We envision hardware developers using the description of the noise to efficiently discover previously unknown properties of devices beyond regimes that have been explored at present. This should enable engineering effects to be concentrated on removing the most relevant residual errors. Moreover, we envision theoretical efforts that will result in error mitigation and correction techniques that are tailored to the specific noise afflicting a device as reconstructed using the methods in this paper. Such techniques will result in substantially better device performance by enabling optimized design of codes and decoders, bespoke fault tolerance schemes, and error-aware compiling for quantum simulation.

While we have proven rigorous upper bounds, many interesting open questions remain. For example, we have not attempted to prove lower bounds on the problems that we study in this work. It would be interesting to prove such bounds or to find procedures with even better asymptotic scaling.

Other natural avenues for future work are to explicitly prove robustness to gate-dependent noise and to generalize the method to estimate the noise on an interleaved Clifford gate [81]. We have also proven our convergence guarantees in the idealized "single-shot" regime of Ref. [82]. It should be straightforward to generalize our arguments to handle the reuse of individual sequences instead of using a fresh random sequence each time. We have also made exclusive use of the *qubit* Pauli group, but it would be interesting to extend this work to d-level quantum systems as well and more formally incorporate techniques from compressed sensing.

For the reconstruction of Pauli channels as Markov random fields, it should also be possible to incorporate structure learning techniques [56–60] to learn the dependency structure. If the correlations are promised to be bounded degree, then we expect that structure learning can be also be done efficiently, in time poly(n) in the number of qubits. The sample complexity can be improved by making use of parallel samples on the same system so long as the measurements being used in parallel lie in disjoint neighborhoods of the factor graph. This can be used to improve the efficiency still further. Finally, a natural generalization of this idea is to learn a matrix product operator quantum channel [83, 84]. The state and unitary analogs of this idea have been explored before [49, 50], but not yet in a way that is robust to SPAM errors. Perhaps combining the ideas in the present paper with those of Kimmel *et al.* [53] could lead to a more general procedure.

Finally, there is the practical application of our methods to real world quantum devices. A first exploration along these lines can be found in a companion paper by R. Harper and the present authors [39].

ACKNOWLEDGMENTS

We thank Robin Harper and Wenjun Yu for comments on an earlier draft. This work was supported in part by the US Army Research Office grant numbers W911NF-14-1-0098 and W911NF-14-1-0103, the Australian Research Council Centre of Excellence for Engineered Quantum Systems (EQUS) CE170100009, the Government of Ontario, and the Government of Canada through the Canada First Research Excellence Fund (CFREF) and Transformative Quantum Technologies (TQT), the Natural Sciences and Engineering Research Council (NSERC), Industry Canada.

[1] Barbara M. Terhal, "Quantum error correction for quantum memories," Rev. Mod. Phys. **87**, 307 (2015), arXiv:1302.3428.

[2] Joel J. Wallman and Joseph Emerson, "Noise tailoring for scalable quantum computation via randomized compiling," Phys. Rev. A **94**, 052325 (2016), arXiv:1512.01098.

[3] Yuval R. Sanders, Joel J. Wallman, and Barry C. Sanders, "Bounding Quantum Gate Error Rate Based on Reported Gate Fidelity," New J. Phys. **18**, 012002 (2015), arXiv:1501.04932.

[4] Richard Kueng, David M. Long, Andrew C. Doherty, and Steven T. Flammia, "Comparing experiments to the fault-tolerance threshold," Phys. Rev. Lett. **117**, 170502 (2016), arXiv:1510.05653.

[5] Eric Huang, Andrew C. Doherty, and Steven Flammia, "Performance of quantum error correction with coherent errors," Phys. Rev. A **99**, 022313 (2019), arXiv:1805.08227.

[6] Stefanie Beale, Joel Wallman, Mauricio Gutiérrez, Kenneth R. Brown, and Raymond Laflamme, "Coherence in quantum error-correcting codes," Phys. Rev. Lett. **121**, 190501 (2018), arXiv:1805.08802.

[7] Matthew Ware, Guilhem Ribeill, Diego Riste, Colm A. Ryan, Blake Johnson, and Marcus P. da Silva, "Experimental demonstration of Pauli-frame randomization on a superconducting qubit," (2018), 1803.01818.

[8] H. Bombin, Ruben S. Andrist, Masayuki Ohzeki, Helmut G. Katzgraber, and M. A. Martin-Delgado, "Strong resilience of topological codes to depolarization," Physical Review X **2**, 021004 (2012), arXiv:1202.1852.

[9] Naomi H Nickerson and Benjamin J Brown, "Analysing correlated noise on the surface code using adaptive decoding algorithms," Quantum **3**, 131 (2019), arXiv:1712.00502.

[10] Andrew S. Darmawan and David Poulin, "Tensor-network simulations of the surface code under realistic noise," Phys. Rev. Lett. **119**, 040502 (2017), arXiv:1607.06460.

[11] Nishad Maskara, Aleksander Kubica, and Tomas Jochym-O'Connor, "Advantages of versatile neural-network decoding for topological codes," , 1–13 (2018), arXiv:1802.08680.

[12] David K. Tuckett, Stephen D. Bartlett, and Steven T. Flammia, "Ultrahigh error threshold for surface codes with biased noise," Phys. Rev. Lett. **120**, 050505 (2018), arXiv:1708.08474.

[13] David K Tuckett, Christopher T Chubb, Sergey Bravyi, Stephen D Bartlett, and Steven T Flammia, "Tailoring surface codes for highly biased noise," (2018), arXiv:1812.08186.

[14] Alan Robertson, Christopher Granade, Stephen D. Bartlett, and Steven T. Flammia, "Tailored codes for small quantum memories," Phys. Rev. Applied **8**, 064004 (2017), arXiv:1703.08179.

[15] Panos Aliferis and John Preskill, "Fault-tolerant quantum computation against biased noise," Phys. Rev. A **78**, 052331 (2008), arXiv:0710.1301.

[16] Christopher T. Chubb and Steven T. Flammia, "Statistical mechanical models for quantum codes with correlated noise," (2018), arXiv:1809.10704.

[17] Joseph Emerson, Robert Alicki, and Karol Życzkowski, "Scalable noise estimation with random unitary operators," J. Opt. B **7**, S347 (2005), quant-ph/0503243.

[18] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine, "Exact and approximate unitary 2-designs and their application to fidelity estimation," Phys. Rev. A **80**, 012304 (2009), arXiv:quant-ph/0606161.

[19] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer,

R. Ozeri, S. Seidelin, and D. J. Wineland, "Randomized benchmarking of quantum gates," Phys. Rev. A **77**, 012307 (2008), arXiv:0707.0963.

[20] Jonas Helsen, Xiao Xue, Lieven MK Vandersypen, and Stephanie Wehner, "A new class of efficient randomized benchmarking protocols," (2018), arXiv:1806.02048.

[21] Alexander Erhard, Joel James Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban Adrian Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt, "Characterizing large-scale quantum computers via cycle benchmarking," arXiv e-prints , arXiv:1902.08543 (2019), arXiv:1902.08543 [quant-ph].

[22] Tobias Chasseur and Frank K. Wilhelm, "Complete randomized benchmarking protocol accounting for leakage errors," Physical Review A **92**, 042333 (2015), arXiv:1505.00580v2.

[23] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout, "What randomized benchmarking actually measures," Phys. Rev. Lett. **119**, 130502 (2017), arXiv:1702.01853.

[24] Joel Wallman, "Randomized benchmarking with gate-dependent noise," Quantum **2**, 47 (2018), arXiv:1703.09835.

[25] Seth T. Merkel, Emily J. Pritchett, and Bryan H. Fong, "Randomized Benchmarking as Convolution: Fourier Analysis of Gate Dependent Errors," (2018), arXiv:1804.05951.

[26] T. P. Harty, D. T. C. Allcock, C. J. Ballance, L. Guidoni, H. A. Janacek, N. M. Linke, D. N. Stacey, and D. M. Lucas, "High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit," Phys. Rev. Lett. **113**, 220501 (2014), arXiv:1403.1524.

[27] Joseph Emerson, Marcus Silva, Osama Moussa, Colm Ryan, Martin Laforest, Jonathan Baugh, David G. Cory, and Raymond Laflamme, "Symmetrized characterization of noisy quantum processes," Science **317**, 1893–1896 (2007), arXiv:0707.0685.

[28] C. A. Ryan, M. Laforest, and R. Laflamme, "Randomized benchmarking of single- and multi-qubit control in liquid-state NMR quantum information processing," New J. Phys. **11**, 013034 (2009), arXiv:0808.3973.

[29] Jay M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, John A. Smolin, Jerry M. Chow, Colm A. Ryan, Chad Rigetti, S. Poletto, Thomas A. Ohki, Mark B. Ketchen, and M. Steffen, "Characterization of addressability by simultaneous randomized benchmarking," Phys. Rev. Lett. **109**, 240504 (2012), arXiv:1204.6308.

[30] R. Barends, J. Kelly, A. Veitia, A. Megrant, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, I.-C. Hoi, E. Jeffrey, C. Neill, P. J. J. O'Malley, J. Mutus, C. Quintana, P. Roushan, D. Sank, J. Wenner, T. C. White, A. N. Korotkov, A. N. Cleland, and John M. Martinis, "Rolling quantum dice with a superconducting qubit," Phys. Rev. A **90**, 030303 (2014), arXiv:1406.3364.

[31] Arnaud Carignan-Dugas, Joel J. Wallman, and Joseph Emerson, "Characterizing universal gate sets via dihedral benchmarking," Phys. Rev. A **92**, 060302 (2015), arXiv:1508.06312.

[32] Andrew W Cross, Easwar Magesan, Lev S Bishop, John A Smolin, and Jay M Gambetta, "Scalable randomised benchmarking of non-clifford gates," npj Quantum Information **2** (2016), 10.1038/npjqi.2016.12, arXiv:1510.02720.

[33] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, "Real Randomized Benchmarking," Quantum **2**, 85 (2018), arXiv:1801.06121.

[34] W. G. Brown and B. Eastin, "Randomized benchmarking with restricted gate sets," Phys. Rev. A **97**, 062323 (2018), arXiv:1801.04042.

[35] D. S. França and A. K. Hashagen, "Approximate randomized benchmarking for finite groups," Journal of Physics A: Mathematical and Theoretical **51**, 395302 (2018), arXiv:1803.03621.

[36] Robin Harper, Ian Hincks, Chris Ferrie, Steven T. Flammia, and Joel J. Wallman, "Statistical analysis of randomized benchmarking," Phys. Rev. A (2019), arXiv:1901.00535.

[37] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak, "Nonexponential fidelity decay in randomized benchmarking with low-frequency noise," Phys. Rev. A **92**, 022326 (2015), arXiv:1502.05119.

[38] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," (1971), available at `http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf`.

[39] R. Harper, S. T. Flammia, and J. J. Wallman, "Efficient learning of quantum noise," In preparation (2019).

[40] Andrey V. Rodionov, Andrzej Veitia, R. Barends, J. Kelly, Daniel Sank, J. Wenner, John M. Martinis, Robert L. Kosut, and Alexander N. Korotkov, "Compressed sensing quantum process tomography for superconducting quantum gates," Phys. Rev. B **90**, 144504 (2014), arXiv:1407.0761.

[41] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert, "Quantum state

tomography via compressed sensing," Phys. Rev. Lett. **105**, 150401 (2010), arXiv:0909.3304.

[42] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity, and efficient estimators," New J. Phys. **14**, 095022 (2012), arXiv:1205.2300.

[43] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi, and A. G. White, "Efficient measurement of quantum dynamics via compressive sensing," Phys. Rev. Lett. **106**, 100401 (2011), arXiv:0910.5498.

[44] Isaac L. Chuang and M. A. Nielsen, "Prescription for experimental determination of the dynamics of a quantum black box," J. Mod. Opt. **44**, 2455–2467 (1997), quant-ph/9610001.

[45] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu, "Sample-optimal tomography of quantum states," IEEE Transactions on Information Theory , 1 (2017), arXiv:1508.01797.

[46] Ryan O'Donnell and John Wright, "Efficient quantum tomography," in *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16 (ACM, New York, NY, USA, 2016) pp. 899–912.

[47] Ryan O'Donnell and John Wright, "Efficient quantum tomography II," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017 (ACM, New York, NY, USA, 2017) pp. 962–974.

[48] Scott Aaronson, "Shadow tomography of quantum states," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018 (ACM, New York, NY, USA, 2018) pp. 325–338, arXiv:1711.01053.

[49] Marcus Cramer, Martin B. Plenio, Steven T. Flammia, Rolando Somma, David Gross, Stephen D. Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu, "Efficient quantum state tomography," Nat Commun **1**, 149 (2010), arXiv:1101.4366.

[50] M. Holzäpfel, T. Baumgratz, M. Cramer, and M. B. Plenio, "Scalable reconstruction of unitary processes and Hamiltonians," Phys. Rev. A **91**, 042129 (2015), arXiv:1411.6379.

[51] Seth T. Merkel, Jay M. Gambetta, John A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, Colm A. Ryan, and M. Steffen, "Self-consistent quantum process tomography," Phys. Rev. A **87**, 062119 (2013), 1211.0322.

[52] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Kenneth Rudinger, Jonathan Mizrahi, Kevin Fortier, and Peter Maunz, "Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography," Nature Communications **8**, (2016), arXiv:1605.07674.

[53] Shelby Kimmel, Marcus P. da Silva, Colm A. Ryan, Blake R. Johnson, and Thomas Ohki, "Robust extraction of tomographic information via randomized benchmarking," Phys. Rev. X **4**, 011050 (2014), arXiv:1306.2348.

[54] Ingo Roth, Richard Kueng, Shelby Kimmel, Yi-Kai Liu, David Gross, Jens Eisert, and Martin Kliesch, "Recovering quantum gates from few average gate fidelities," Phys. Rev. Lett. **121**, 170502 (2018), arXiv:1803.00572.

[55] Guy Bresler, David Gamarnik, and Devavrat Shah, "Hardness of parameter estimation in graphical models," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14 (MIT Press, Cambridge, MA, USA, 2014) pp. 1062–1070, arXiv:1409.3836.

[56] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," IEEE Transactions on Information Theory **14**, 462–467 (1968).

[57] Pieter Abbeel, Daphne Koller, and Andrew Y. Ng, "Learning factor graphs in polynomial time and sample complexity," Journal of Machine Learning Research **7**, 1743 (2006), arXiv:1207.1366.

[58] Guy Bresler, Elchanan Mossel, and Allan Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," SIAM Journal on Computing **42**, 563–578 (2013), arXiv:0712.1402.

[59] Guy Bresler, "Efficiently learning Ising models on arbitrary graphs," in *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15 (ACM, New York, NY, USA, 2015) pp. 771–782, arXiv:1411.6156.

[60] Linus Hamilton, Frederic Koehler, and Ankur Moitra, "Information theoretic properties of Markov random fields, and their algorithmic applications," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Curran Associates Inc., USA, 2017) pp. 2460–2469, arXiv:1705.11107.

[61] A. Klivans and R. Meka, "Learning graphical models using multiplicative weights," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017) pp. 343–354, arXiv:1706.06274.

[62] Daniel Gottesman, *Stabilizer codes and quantum error correction*, Ph.D. thesis, Caltech, Pasadena, California (1997), quant-ph/9705052.

[63] W. K. Wootters and B. D. Fields, "Optimal state-determination by mutually unbiased measurements," Annals of Physics **191**, 363 (1989).

[64] Easwar Magesan, *Characterizing Noise in Quantum Systems*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada (2012).

[65] Timothy J Proctor, Arnaud Carignan-Dugas, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young, "Direct randomized benchmarking for multi-qubit devices," (2018), arXiv:1807.07975.

[66] Wassily Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," Journal of the American Statistical Association **58**, 13 (1963).

[67] R. J. Serfling, "Probability Inequalities for the Sum in Sampling without Replacement," The Annals of Statistics **2**, 39 (1974).

[68] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" IEEE Trans. Info. Theory **52**, 5406–5425 (2006).

[69] E Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inform. Theory **52**, 489 (2006).

[70] D.L. Donoho, "Compressed sensing," IEEE Trans. Info. Theory **52**, 1289–1306 (2006).

[71] Robin Scheibler, Saeid Haghighatshoar, and Martin Vetterli, "A fast Hadamard transform for signals with sublinear sparsity in the transform domain," IEEE Transactions on Information Theory **61**, 2115–2132 (2015), arXiv:1310.1803.

[72] Mahdi Cheraghchi and Piotr Indyk, "Nearly optimal deterministic algorithm for sparse Walsh-Hadamard transform," ACM Trans. Algorithms **13**, 34:1–34:36 (2017), arXiv:1504.07648.

[73] Xiao Li, Joseph K. Bradley, Sameer Pawar, and Kannan Ramchandran, "SPRIGHT: A fast and robust framework for sparse Walsh-Hadamard transform," (2015), arXiv:1508.06336.

[74] Yi Janet Lu, "Walsh sampling with incomplete noisy signals," in *Advances in Intelligent Systems and Computing* (Springer International Publishing, 2018) pp. 131–144, arXiv:1602.00095.

[75] Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques* (MIT Press, Cambridge, MA, 2009).

[76] Julian Besag, "Spatial interaction and the statistical analysis of lattice systems," Journal of the Royal Statistical Society. Series B (Methodological) **36**, 192–236 (1974).

[77] F Barahona, "On the computational complexity of Ising spin glass models," Journal of Physics A: Mathematical and General **15**, 3241 (1982).

[78] Mark Jerrum and Alistair Sinclair, "Polynomial-time approximation algorithms for the Ising model," SIAM Journal on Computing **22**, 1087–1116 (1993).

[79] M.S. Pinsker, *Information and information stability of random variables and processes* (Holden-Day, 1964).

[80] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory* (Wiley, 1991).

[81] Easwar Magesan, Jay M. Gambetta, B. R. Johnson, Colm A. Ryan, Jerry M. Chow, Seth T. Merkel, Marcus P. da Silva, George A. Keefe, Mary B. Rothwell, Thomas A. Ohki, Mark B. Ketchen, and M. Steffen, "Efficient measurement of quantum gate error by interleaved randomized benchmarking," Phys. Rev. Lett. **109**, 080505 (2012), arXiv:1203.4550.

[82] Christopher Granade, Christopher Ferrie, and D. G. Cory, "Accelerated Randomized Benchmarking," New J. Phys. **17**, 013042 (2015), arXiv:1404.5275.

[83] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac, "Matrix product density operators: Simulation of finite-temperature and dissipative systems," Phys. Rev. Lett. **93**, 207204 (2004), cond-mat/0406426.

[84] Michael Zwolak and Guifré Vidal, "Mixed-state dynamics in one-dimensional quantum lattice systems: A time-dependent superoperator renormalization algorithm," Phys. Rev. Lett. **93**, 207205 (2004), cond-mat/0406440.