Abstract

With the advent of quantum processors exceeding 100 qubits and the high engineering complexities involved, there is a need for holistically benchmarking the processor to have quality assurance. Linear cross-entropy benchmarking (XEB) has been used extensively for systems with 50 or more qubits but is fundamentally limited in scale due to the exponentially large computational resources required for classical simulation. In this work we propose conducting linear XEB with Clifford circuits, a scheme we call Clifford XEB. Since Clifford circuits can be simulated in polynomial time, Clifford XEB can be scaled to much larger systems. To validate this claim, we run numerical simulations for particular classes of Clifford circuits with noise and observe exponential decays. When noise levels are low, the decay rates are well-correlated with the noise of each cycle assuming a digital error model. We perform simulations of systems up to 1,225 qubits, where the classical processing task can be easily dealt with by a workstation. Furthermore, using the theoretical guarantees in Chen et al. (arXiv:2203.12703), we prove that Clifford XEB with our proposed Clifford circuits must yield exponential decays under a general error model for sufficiently low errors. Our theoretical results explain some of the phenomena observed in the simulations and shed light on the behavior of general linear XEB experiments.
1 Introduction

Quantum computers are no longer theoretical constructs or even small-scale, proof-of-concept devices. There are already multiple processors with over 50 qubits [1, 2], and recently a processor with more than 100 qubits has been built [3]. With this number of qubits comes a litany of engineering and operating challenges, including the overwhelming number of control lines and pulse generators required to perform gates and measurements. There are also challenges in processor design, including significant crosstalk between qubits and poor two-qubit gate connectivity. Given all these challenges, there is a demand for benchmarking these processors to have assurance that the algorithms we run will give accurate results. In addition, the figure of merit provided by such a benchmarking scheme can serve as an essential guide for hardware iteration to build ever higher quality processors. Although common methods such as single- and two-qubit randomized benchmarking (RB) can provide useful information, these methods are not sensitive to aggregate effects such as the difference between calibrating qubits on a processor individually or simultaneously [4]. To capture such system-level effects, holistic benchmarks which should involve all or a large number of qubits on the processor are needed. Such benchmarks can also provide information about whether there are correlated errors in the system by comparing the results to the digital error model [1]. This is crucial for realizing fault-tolerant quantum computation.

To holistically benchmark modern quantum processors with tens of qubits, various schemes have been developed, most notably linear cross-entropy benchmarking (linear XEB). Linear XEB was originally proposed for the “quantum supremacy” experiment [1], where it was used to characterize increasingly larger quantum circuits so as to extrapolate the error of the 20-cycle Sycamore circuit. To implement linear XEB, we run quantum circuits with random layers of gates and perform a computational basis measurement. We next classically compute the probability of the bit string we measure. We then repeat, taking an overall average. This average, up to constants, is the linear XEB measure. It has been experimentally and numerically observed that this measure exponentially decays with the number of cycles for a noisy circuit, and this decay exponent is proposed as a measure of gate quality [1, 2, 5]. Although originally conceived to support the “quantum supremacy” claim, linear XEB has become a benchmarking scheme in its own right [1, 2, 6, 7]. Linear XEB has the advantage of requiring only a shallow circuit, which is easy to implement on current processors. However, if we want to benchmark larger devices, linear XEB has an obvious Achilles’ heel. Namely, the random gate set is chosen such that the classical processing needed for linear XEB, that is, simulating the resulting random quantum circuits, scales poorly with the number of qubits. Indeed, by the quantum supremacy claim itself, linear XEB by design cannot benchmark a processor with
more than a few tens of qubits [1]. Therefore, if we want to keep apace with current hardware development, we need an alternative benchmarking scheme that can scale to larger qubit numbers.

In this work, we propose a scalable benchmarking scheme which replaces the random circuits in linear XEB with Clifford circuits. We call our scheme Clifford XEB. This is a crucial modification since Clifford circuits are easy to classically simulate [8]. Moreover, Clifford XEB retains the convenience of requiring only shallow circuits. We believe that given the prevalence of Clifford RB [9, 10, 11] and linear XEB, it is a negligible overhead to adopt Clifford XEB as a new benchmarking scheme. To demonstrate the scalability of our scheme, we perform numerical simulations of certain random Clifford circuits on various topologies with depolarizing noise. We simulate circuits with up to 225 qubits. For noise levels comparable to what is possible in current hardware, we can see clear exponential decays whose rates are consistent with the digital error model. We also conduct simulations on the 2-D grid topology with 1,225 qubits to further showcase the scalability of Clifford XEB. On top of the promising numerical results, we also prove that Clifford XEB for the classes of circuits we consider yields an exponential decay under a general error model for sufficiently low error. This is necessary since we a priori do not know if we would even measure an exponential decay for every experimental setting we encounter (although this can be shown for special error models [7, 5]). In addition to providing this assurance, our proof also explains some of the phenomena we observe in the simulations, such as the linear XEB measure rapidly decaying for low cycle numbers before entering a smooth exponential decay. Through our theoretical results we provide concrete guidelines for how experiments, both Clifford XEB and conventional linear XEB, should be interpreted so as to correctly extract information about gate errors.

2 The Clifford XEB Scheme

In this section we give an explicit description of the Clifford XEB scheme and perform numerical simulations showing its viability for benchmarking quantum processors with more than a thousand qubits.

2.1 Description of the Scheme

Consider a quantum computer with \( n \) qubits. Let \( S \subset SU(2^n) \) be a subset of the special unitary group and let \( \mu \) be a probability distribution on \( S \). The implementation map \( \phi : S \rightarrow \mathcal{C}(2^n) \), where \( \mathcal{C}(2^n) \) is the set of quantum channels on \( n \) qubits, defines the noisy implementation of gates on the system. This in particular means that the noise for each Clifford element is independent of the previous elements, that is, the noise is Markovian. Assume the initial state is \( \rho_0 \) and final measurement POVM is \( M = \{ M_x \}_{x \in \{0,1\}^n} \), which are noisy realizations of some desired initial state and measurement. Then, a linear XEB experiment is defined as follows.

Definition 1 (Linear XEB). A linear XEB scheme with parameters \((S, \mu, \phi, M, \rho_0)\) is given by the following procedure:

1. For a given number of cycles \( m \), sample \( m \) i.i.d. quantum gates \( g_1, \ldots, g_m \) from \( S \) according to the distribution \( \mu \).

---

\(^1\)Readers who are familiar with the “quantum supremacy” paper may point out that the paper’s method of using patch circuits or elided circuits can push the qubit number further. However, using patch circuits is effectively benchmarking two separate processors. As for elided circuits, the complexity of classical simulation still scales poorly.
2. Initialize the system to $\rho_0$ and apply $\phi(g_1), \phi(g_2), \ldots, \phi(g_m)$ to the system.

3. Measure the system under a POVM $\{M_x\}_{x \in \{0,1\}^n}$ and get a binary string $x$. Calculate the probability of getting $x$ for the ideal circuit: $p_x(m) = |\langle x | g_m \ldots g_2 g_1 | 0^n \rangle|^2$.

4. Repeat steps 2 to 3 multiple times for this sampled circuit and calculate the average of $p_x(m)$ over sampled strings $x$.

5. Repeat steps 1 to 4 multiple times and let $\hat{p}(m)$ be the average of $\mathbb{E}_x[p_x(m)]$ over different circuits. Let $\hat{q}(m) = -1 + 2^n \hat{p}(m)$.

6. Repeat steps 1 to 5 for number of cycles $m_1, \ldots, m_k$. The returned results are $\hat{q}(m_1), \ldots, \hat{q}(m_k)$.

We can see that $\hat{q}(m)$ is an unbiased estimator of the following quantity

$$q_R(m) = -1 + 2^n \mathbb{E}_{g_1,\ldots,g_m \sim \mu} \left[ \sum_{x \in \{0,1\}^n} |\langle x | g_m \ldots g_2 g_1 | 0^n \rangle|^2 \tilde{p}(x) \right],$$

where $\tilde{p}(x)$ is the probability of measuring the binary string $x$ in the noisy circuit:

$$\tilde{p}(x) = \text{tr}[M_x \cdot \phi(g_m) \circ \cdots \circ \phi(g_2) \circ \phi(g_1)(\rho_0)].$$

We next define Clifford XEB, a special case, illustrated in Figure 1.

**Definition 2** (Clifford XEB). A Clifford XEB scheme with parameters $(S, \mu, \phi, M, \rho_0)$ is a linear XEB scheme, where $S \subseteq \text{Cl}(2^n)$ is a subset of the Clifford group.

![Figure 1: Illustration of one iteration of a Clifford XEB experiment. We start with an initial state, apply $m$ cycles of noisy implementations of Clifford elements, and finally perform a measurement. The measured bit string’s ideal probability is then classically computed. This value is averaged over multiple measurements and random circuits.](image)

That is, this Clifford XEB only differs from conventional linear XEB in that we require $S$ to be Clifford. Again, we do this since the ideal distributions of Clifford circuits can be efficiently computed classically [8]. Also note that our definition of cycles is slightly different from that of [1]. There they allow heterogeneous circuit geometries for every cycle, while we require that the gates in each cycle is sampled from the same distribution.

### 2.2 Numerical Simulations

We numerically study the behavior of Clifford XEB for various connectivity topologies to demonstrate its wide applicability. Our findings also help explain the behavior of general linear XEB schemes. The topologies we consider include a 1-D chain, a 2-D grid, and a star topology, i.e. there is a center qubit that is connected to all the other qubits, to implement the construction in [12], which we will refer to as the Clifford approximate twirl. We describe the particular distribution over Cliffords for each topology, illustrated in Figure 2.
1. For the 1-D chain, each cycle consists of two layers of two-qubit gates each preceded by a layer of single qubit gates. The single-qubit gate layer involves i.i.d. random single-qubit Clifford gates on each qubit. The two-qubit layers are CNOT gates: the first connecting qubits (0, 1), (2, 3), · · · , (2k, 2k+1), · · · and the second connecting (1, 2), (3, 4), · · · , (2k+1, 2k+2), · · · .

2. For the 2-D grid, we take motivation from the random circuits on the Sycamore processor in [1]. Each cycle consists of a single-qubit gate layer consisting of i.i.d. single-qubit Cliffords on every qubit, followed by a two-qubit gate layer consisting of CNOT gates. The structure of the CNOT’s is chosen from the configurations A, B, C and D in [1, Figure 3] with equal probability.\(^2\)

3. We implement the Clifford approximate twirl [12] on a star topology. Each cycle in this construction consists of two repeated circuits, each circuit involving three layers of two-qubit gates, interleaved by four layers of single-qubit gates. See Appendix A for the detailed construction.

The numerical experiments are run using the stabilizer-based Clifford circuit simulator Stim [13] on a 40-core Intel Xeon Platinum machine with 96GB memory. For each topology, we run experiments with different qubit numbers and noise levels. We pick qubit numbers 25, 100 and 225, which are respectively 5 × 5, 10 × 10, and 15 × 15 on the 2-D grid. As for noise levels, we assume depolarizing noise for single- and two-qubit gates of (10\(^{-5}\), 10\(^{-4}\)), (10\(^{-4}\), 10\(^{-3}\)), and (10\(^{-3}\), 10\(^{-2}\)) respectively. This is similar to what is achievable in current hardware. We also simulate the ideal case for reference. Each experiment is run for 2 to 50 cycles, where for each cycle number 3,000 random circuits are generated, each sampled 100,000 times. The results are plotted in Figure 3. All of the experiments added together only take a total of several hours on the aforementioned workstation. In comparison, conventional linear XEB is already infeasible for just tens of qubits.

From Figure 3 we can make the following observations.

**Convergence and two-phase behavior** We see a universal behavior of linear XEB with increasing cycle number: the ideal value always converges to 1,\(^3\) while the noisy value converges to 0. This is the same as the experimental findings in [1]. Intuitively, for sufficiently many cycles, the

---

\(^2\)In the Sycamore random circuit sampling experiments, the two-qubit gates in different cycles follow a fixed pattern ABCDCDAB. We do not do this because we require every cycle to have the same distribution.

\(^3\)Technically, the ideal linear XEB value converges to \(2^{-1}\) as shown in Appendix B.1, \(D\) being the dimension of the quantum system. In the rest of the paper we take the approximation \(2^{-1} \approx 1\) unless explicitly stated.
The $x$-axis is twice the cycle number since we need to apply the circuit construction twice to get a $\gamma$-approximate twirl. See Section 3 for details.

Figure 3: Linear XEB as a function of the number of cycles on different topologies, qubit numbers, and noise levels. Each row corresponds to the labeled topology. The number of qubits simulated on each topology is, from left to right, 25, 100, and 225. The horizontal line is linear XEB $= 1$ for reference.
Figure 4: Comparison between the linear XEB decay rate, the digital error model, and the mixing rate, under different topologies. The blue dashed line depicts the decay rate of the ideal linear XEB towards 1. For each error rate, the solid lines are predictions given by the digital error model, and the dots are extracted from the numerical experiments. Some data is omitted due to failure of fitting.

random Clifford circuit will converge to a uniform distribution over the Clifford group [14], while a noisy implementation will eventually lose all information of the input. This explains the final convergence values. A more rigorous proof is given in Appendix B.1 and Appendix B.2.

The curves also exhibit a two-phase behavior: In the first phase, which we call the **scrambling phase** in reference to the connection between random circuits and quantum scrambling [15, 16, 17, 18, 19, 14, 20, 21], both the ideal and noisy linear XEB values decay quickly from being exponentially large to a constant value. The second phase, which we call the **decaying phase**, the ideal linear XEB value converges to 1, while the noisy values undergo single exponential decays (which would be a linear decay on the semilog plot). Intuitively, only the data in the second phase is of relevance to benchmarking, as the first phase is characterized by the **scrambling** of quantum information, rather than the **loss** of it. We will give theoretical support for this intuition in Section 3.

One way of interpreting the decay rate is to use it as a proxy for the fidelity of a cycle, which in turn can be estimated using the digital error model [1] by multiplying the fidelity of each gate in the circuit. To see if this applies to the numerical experiments, we extract the decay rates by linearly fitting the curve in the second phase (we manually choose the range of cycles). We then compare them with the digital error model in Figure 4. We see that the digital error model succeeds to predict the linear XEB decay rate when the error rate is small, but fails to do so when the error rate is too large. This is because even in the decaying phase, the ideal linear XEB term contributes an error term on top of the exponential decay from gate errors. As a result, when the gate error exceeds a certain threshold, the decay of the linear XEB value no longer represents the fidelity but is dominated by the slower mixing of the ideal term.

To verify this intuition, for each circuit configuration (topology and number of qubits) we sampled 300,000 random circuits to better evaluate the mixing rate of the ideal linear XEB value, which we call the **mixing rate**. We do this because of the quick decay of the noiseless value and the high variance of Clifford XEB (See Appendix B.1). The mixing rate towards 1 is plotted as dashed lines in Figure 4. It can be observed that the digital error model fails to predict the linear XEB decay rate when the fidelity estimated from the digital error model is below the line of mixing of the ideal linear XEB. This indicates that the Clifford XEB, and in general linear XEB, only provides useful information when the noise level is sufficiently small. As a rule of thumb, one can estimate
the convergence rate of the ideal linear XEB values prior to running linear XEB experiments to ensure that the extracted decay rate from the experiment reflects the gate fidelities. To characterize large errors, one could either change the twirling distribution $\mu$ for faster scrambling (e.g. the one given in Appendix A gives a theoretically provable mixing rate) or benchmark a subset of the qubits on the device to lower the error of each cycle.

**Fluctuations** Occasionally (e.g. in the rightmost plot of Figure 3c), there are noticeable bumps in the curves. Moreover we observe that the red line, corresponding to a two-qubit gate error rate of $10^{-2}$, fluctuates heavily when the linear XEB value is small. We attribute such instabilities to three main causes:

- During the scrambling phase, both the expectation and the variance of the noiseless and noisy linear XEB value is large, resulting in significant fluctuations for the noiseless and noisy curves. The middle plot in Figure 3b has such a bump at $m = 14$.

- Unlike non-Clifford scrambling circuits that almost surely results in a measurement distribution obeying Porter-Thomas statistics, the measurement distribution of a Clifford circuit is always uniform on the support. As a result, the noiseless linear XEB value has constant variance even after the scrambling phase. This is different from doing linear XEB with Haar random gates because the Clifford group forms a unitary three-design but not a four-design. We derive the variance in Appendix B.1. Since the variance is constant, taking a large but constant number of random circuits smooths out the fluctuation. It can be observed in our plot that taking 3000 random circuits per data points effectively resolves this issue, except for a large deviation observed in the rightmost plot of Figure 3c, at 18 cycles.

- The linear XEB values of the noisy curves converge to 0 exponentially with respect to the cycle number. Given a fixed number of samples per random circuit, the absolute error of the linear XEB value is a constant, but the relative error quickly grows as the expectation approaches 0. Similar to [1], sufficiently many random samples needs to be taken from each random circuit to ensure an accurate evaluation of the expectation value for linear XEB. This can be observed in the right two figures in Figure 3b where the red lines fluctuate at the tails, and in Figure 3c where the red lines are not even observed in our range.

Understanding these sources of instability can help guide Clifford XEB and general linear XEB experiments. The second source only applies to Clifford XEB, but a constant number of random circuits suffices to reduce the variance to a small number. The third source applies for both Clifford XEB and linear XEB, in fact for any RB-based scheme. The number of samples taken must be sufficiently large such that the exponential decay curves can be confidently recovered. However, note that in our simulations we can often already isolate a clean single exponential decay at not too small linear XEB values for which we do not need as many samples.

To accurately extract decay rates, we need to perform Clifford XEB experiments in the decaying phase rather than the scrambling phase. A natural question to ask is at what cycle number a Clifford XEB scheme enters the decaying phase, i.e. the scrambling time. While we give evidence in Section 3 and Appendix B.2 that the scrambling time could scale linearly with respect to number of qubits, we observe experimentally that this only applies to Clifford approximate twist (scrambling times are approximately 4, 9, 18 for 25, 100, 225 qubits, respectively). Scrambling happens much faster for the 1-D chain and 2-D grid experiments and seems only weakly dependent on the number of qubits.
Although a finer theoretical analysis of the scrambling behavior of the noiseless experiment is needed to fully explain this phenomenon, we note that classical simulation of the noiseless experiment can be used to determine both the scrambling time and the mixing rate in the decaying phase. Such a simulation can be done prior to any physical experiment and serves as a guideline for the depths of the actual circuits to be run.

Finally, we present the results of Clifford XEB on the 2-D grid with $35 \times 35 = 1225$ qubits, the results shown in Figure 5. The experiments are run with two-qubit error rates ranging from $10^{-5}$ to $10^{-3}$ due to the larger number of qubits. We plot linear XEB for cycle numbers from 20 to 100 because of longer scrambling time. We see that, while the $10^{-3}$ error decay curve is dominated by the mixing rate, Clifford XEB is consistent with the digital error model when the error rate is below the mixing rate. This shows that the Clifford XEB scheme can be readily applied to quantum devices with more than 1,000 qubits. On the aforementioned workstation, all the noisy experiments took approximately 12 hours to run, and two days and a half to run the noiseless experiments in order to reliably extract the mixing rate.

3 The Theory of Clifford XEB

In this section we mathematically prove that Clifford XEB with the 1-D chain distribution, 2-D grid distribution, or the Clifford approximate twirl yields a single exponential decay under a general error model for sufficiently small errors. For the Clifford approximate twirl, this decay is theoretically guaranteed for a qubit number that scales inversely with the infidelity. Our results also support our explanations of phenomena observed in Section 2.2. These technical proofs can be skipped with little effect on the reading of the rest of the paper.

In [22], a generalized framework for RB called universal randomized benchmarking (URB) is proposed, where they prove that for a particular class of schemes known as twirling schemes, an experiment would yield a single exponential decay. Here we show that Clifford XEB with our specific distributions are twirling schemes.

For any $g \in SU(2^n)$, define $\omega(g)$ as the corresponding quantum channel $\omega(g) : \rho \to g\rho g^\dagger$. 

Figure 5: Clifford XEB on a 2-D grid of $35 \times 35 = 1225$ qubits. Data points for curves with noise are evaluated over 3000 random circuits with 100,000 samples each. Data points for the noiseless curve is evaluated over 300,000 random circuits. Predictions given by the digital error model is plotted in dashed dotted lines for comparison.
We first define a $\gamma$-approximate twirl, which is one possible definition of an approximate unitary 2-design [12, 22]:

**Definition 3.** Let $\mu$ be a measure on the unitary group and $\mathcal{C}$ be a linear operator on the space of Hermitian matrices. The twirling map

$$\Lambda(\mu) : \mathcal{C} \mapsto \int_{g \sim \mu} dg \omega(g^\dagger) \circ \mathcal{C} \circ \omega(g)$$

is a $\gamma$-approximate twirl if it satisfies

$$|||\Lambda(\mu) - \Lambda(\mu_H)|||_\diamond \leq \gamma,$$

where $\mu_H$ is the Haar measure and $||| \cdot |||_\diamond$ is the induced diamond norm:

$$|||\Lambda|||_\diamond := \max_C \frac{||\Lambda(C)||_\diamond}{||C||_\diamond}.$$ (3.3)

Alternatively, $\Lambda(\mu)$ is a $\gamma$-approximate twirl in spectral norm if it satisfies

$$|||\Lambda(\mu) - \Lambda(\mu_H)|||_2 \leq \gamma,$$ (3.4)

where $||| \cdot |||_2$ is the spectral norm of the twirling map treated as a linear operator on superoperators.

### 3.1 Clifford Approximate Twirl

#### 3.1.1 Proof of Exponential Decay

For Clifford approximate twirls, we use the following theorem in [22]:

**Theorem 4** (Theorem 8 from [22] and Appendix B.2). Let $(S, \mu, \phi, M, \rho_0)$ be a linear XEB scheme on $n$ qubits. Suppose that the twirling map corresponding to $\mu$ is a $\gamma$-approximate twirl and that the implementation map satisfies

$$E_{g \sim \mu} \|\phi(g) - \omega(g)\|_\diamond \leq \delta.$$ (3.5)

If $\delta \leq \frac{1 - \gamma}{\pi^2}$, then there exists $A, B \in \mathbb{R}$ and $p \in [1 - 2\delta, 1]$, such that

$$|q_R(m) - (A + Bp^m)| \leq 16 \times 2^n(\gamma + 6\delta)^m.$$ (3.6)

Further, when $p < 1$ we have $A = 0$ and $B = O(1)$.

Note that we give here a general result for linear XEB, that is, $S$ need not be Clifford. The main difference from the statement in [22] is the factor of $2^n$ in Equation (3.6). This arises from the factor of $2^n$ in the definition of the XEB quantity in Equation (2.1) which is used for normalization [1]. However, for large $n$ this is problematic since this would make the error term on the RHS in Equation (3.6) unacceptably large, considering that $B = O(1)$. To make the error term smaller than the exponential decay $A + Bp^m$ we want to extract, we need the number of cycles $m$ to scale linearly with $n$: $m = \Omega(n).$\(^4\) Indeed, we do observe that the scrambling time (end of the scrambling phase) is approximately a linear function of the number of qubits in Figure 3c.

\(^4\)Note that Equation (3.6) holds for arbitrarily large $m$. Intuitively, the signal does not disappear at such large $m$ because we assume the noise between each cycle is Markovian, although error correlations are allowed to exist within a cycle.
Now, we obtained the Clifford approximate twirl construction from [12], where they proved that it constitutes a $\gamma$-approximate twirl. Hence, we can directly apply Theorem 4 to Clifford XEB with the Clifford approximate twirl to theoretically guarantee an exponential decay. This guarantee holds under a general error model (we only assume Markovianity), giving much more confidence to what we can expect from running benchmarking experiments. A priori, for schemes where a particular error model is assumed or where there is no theoretical guarantee whatsoever, we might run an experiment and not even see an exponential decay, rendering the scheme completely inapplicable for benchmarking the device in question.

The Clifford approximate twirl is a theoretical construction with excellent circuit size and depth scaling, involving a probabilistic circuit of size $O(n \log 1/\gamma)$ and depth $O(\log n \log 1/\gamma)$. We prove in Appendix C that this is actually optimal in $n$ by showing any $\gamma$-approximate twirl requires a circuit of size $\Omega((1-\gamma)n)$ and depth $\Omega((1-\gamma) \log n)$. However, to achieve this scaling, we would need a complete graph topology. Fortunately, we can easily adapt it to other topologies that are more relevant to hardware at the cost of sub-optimal scaling. For a detailed analysis, see Appendix A.2.

### 3.1.2 Qubit Scaling with Error

Assuming a fixed gate error, any benchmarking scheme has a limit on the number of qubits it can be applied to. One reason for this simply follows from the size of the quantum benchmarking circuit required. To obtain a meaningful result from the circuit, the total error of the circuit should be less than unity.

We consider the error of one cycle of a RB scheme. Now, we are working under a general error model where gate errors within a cycle can be arbitrarily correlated, so we need to consider worst-case additive\(^5\) error accumulation where errors of individual gates add. Then, we must require

\[
s \epsilon \lesssim 1,
\]

where $s$ is the benchmarking circuit size and $\epsilon$ is the individual gate error (assumed to be the same for all gates). Since $s$ is a function of $n$, this gives an upper bound on $n$. In particular, for Clifford RB or any scheme where at least one uniformly random $n$-qubit Clifford element needs to be generated, $s = \Omega(n^2)$, and so

\[
n \lesssim \frac{1}{\sqrt{\epsilon}}.
\]

This is somewhat unsatisfactory; for instance, we would have to lower gate errors by a factor of 100 to be able to benchmark 10 times more qubits.

We show in this section that Clifford XEB with the Clifford approximate twirl can do quadratically better: an inverse linear scaling $n \sim \epsilon^{-1}$. That is, by halving the gate error, we can benchmark twice as many qubits. This is actually optimal for twirling schemes which must implement $\gamma$-approximate twirls given the circuit lower bound we prove in Appendix C. In fact, as long as we have local gates (acts on a constant number of qubits), any benchmarking circuit in which all of the qubits are involved in a gate must have linear size. Thus, under a general error model, the best we can hope for in such a scenario is inverse linear scaling.

\(^5\)If errors are uncorrelated, we expect a multiplicative accumulation where fidelities of individual gates multiply, i.e. the digital error model.
We now show achievability. The Clifford approximate twirl is able to form a γ-approximate twirl for γ ≈ 1/2 with a linear size circuit [12], so the bound on n from that would be $n \lesssim \epsilon^{-1}$. However, we have an additional constraint. In Theorem 4, we require
\[ \delta \leq \frac{1 - \gamma}{11}. \]  
(3.9)
This is a similar requirement to that of the main result in [11]. Intuitively, γ is the distance of our ideal twirling map from the Haar twirling map, which is known to give an exponential decay [12]. Hence, the higher γ is, the lower the tolerance the exponential decay has to the error δ, which is effectively a perturbation on the ideal twirling map. We analyze what Equation (3.9) implies about how the number of qubits we can benchmark scales with gate error for our scheme.

Recall δ is the error of an entire γ-approximate twirl, not just individual gate errors. We assume a simplified error model where single-qubit gates are noiseless and two-qubit gates have the same error rate. Under additive error accumulation, δ is of the form
\[ \delta = s \epsilon, \]  
(3.10)
where $\epsilon$ is the two-qubit gate error and $s$ is the size of the circuit. Now, for the Clifford approximate twirl, $s = cnk$, where $c$ is a constant that appears in the circuit construction, and $k = O(\log(1/\gamma)) \in \mathbb{N}$ is the number of times the circuit is repeated to achieve a γ-approximate twirl [12]. Explicitly, with $k$ repetitions, we achieve a γ-approximate twirl with
\[ \gamma = \frac{1}{2^{k-1}} \left(1 + \frac{1}{4^n}\right)^k. \]  
(3.11)
Hence, Equation (3.9) becomes
\[ cnk \epsilon \leq \frac{1 - \frac{1}{2^{k-1}} \left(1 + \frac{1}{4^n}\right)^k}{11}. \]  
(3.12)
We want to look at the asymptotic case where $n$ is large, so we make a simplifying approximation $\frac{1}{4^n} \approx 0$. Then, Equation (3.12) becomes an upper bound on $n$:
\[ n \leq \frac{1 - 2^{-(k-1)}}{11cke}. \]  
(3.13)
We see that for fixed $\epsilon$, to maximize the upper bound, we set $k = 2$ or $k = 3$ (they give the same answer) to obtain
\[ n \leq \frac{1}{44\epsilon}. \]  
(3.14)
Thus, the qubit number scales inversely with the infidelity.

To be extremely concrete, we can fill in some numbers. State-of-the-art ion trap and superconducting circuit devices can achieve a two-qubit gate infidelity of around $10^{-3}$ [23, 24]. Tracking the constants in [12], $c = 3$ for a star topology circuit. We therefore have a theoretical guarantee of an exponential decay for up to $n = 7$ qubits\(^7\) under a general error model. Note that there was little
\footnote{This is for the sake of numerical convenience and not a limiting assumption on what errors we allow.}
\footnote{That is, $n = 7, k = 2, c = 3, \epsilon = 10^{-3}$ satisfies Equation (3.12).}
attempt in [22] to optimize the constant in Equation (3.9) or in [12] to optimize the circuit size constants, so the largest \( n \) for which we have a theoretical guarantee may be substantially higher. We leave this for future work.

We compare our theoretical predictions with the findings in our numerical simulations. For the Clifford approximate twirl simulations, we only have a theoretical guarantee for the case with 25 qubits and two-qubit gate error rate \( 10^{-4} \) (yellow line in the leftmost plot in Figure 3c) if we fill in all the numbers and the constants. Nevertheless, most of the experiments give rise to single exponential decays that closely reflect the overall noise level of the circuits, which indicates that Clifford XEB is applicable to a wider variety of settings than might be suggested by theory. One reason for this disparity is that Theorem 4 applies to general error models within each cycle, which can be highly correlated and, in the worst case, even adversarial. In contrast, in our numerical experiments we have gate-independent depolarizing errors. Since our scheme is optimal in scaling for arbitrary errors, it is unlikely that a theoretical guarantee for a general error model can be extended to over a hundred qubits given the two-qubit noise levels on current devices. To extend Theorem 4 to larger quantum devices, we either have to reduce gate errors or make assumptions on the error models. For example, it is proven in [25] that the number of qubits can scale with \( \epsilon^{-2} \), provided that the noise level is sufficiently weak and incoherent.\(^8\) In general, we should seek a balance of theoretical guarantee and applicability, i.e., make a minimal number of realistic assumptions about errors to obtain a theoretical guarantee.

### 3.2 1-D Chain and 2-D Grid

We would hope that the theoretical analysis for the Clifford approximate twirl can extend to the schemes on the other topologies we consider. However, for the 1-D chain and 2-D grid schemes, the depth of each circuit cycle is constant. As shown in Appendix C, logarithmic depth is necessary to get an approximate twirl with \( \gamma \) bounded away from 1 by a constant. For these schemes, we instead argue that the measures on the unitary group they induce are \( \gamma \)-approximate twirls with respect to the \( ||| \cdot ||| \) norm. In this case we can use the following theorem proved in [22]:

**Theorem 5** (Corollary 16 of [22]). Let \( (S, \mu, \phi, M, \rho_0) \) be a linear XEB scheme on \( n \) qubits. Suppose that the twirling map corresponding to \( \mu \) is a \( \gamma \)-approximate twirl with respect to the \( ||| \cdot ||| \) norm and that the implementation map satisfies

\[
\mathbb{E}_{g \sim \mu} \| \phi(g) - \omega(g) \|_\diamond \leq \delta. \tag{3.15}
\]

Let \( \delta' := 2^{n/2} \delta \). If \( \delta' \leq \frac{1}{11} \), then there exists \( A, B \in \mathbb{R} \) and \( p \in [1 - 2\delta', 1] \), such that

\[
|q_R(m) - (A + Bp^m)| \leq 16 \times 2^5 \gamma^m (\gamma + 6\delta')^m. \tag{3.16}
\]

Although the \( 2^n \) factors can be very large, this still shows that for sufficiently small errors, we can prove a single exponential decay under a general error model.

In the case that the error channels are mixtures of unitaries, we have another version of Theorem 5 which relaxes the noise level restriction to essentially that of Theorem 4. We have the following.

\(^8\)Note however they considered a setting different from our quest for exponential decays.
Theorem 6. Let \((S, \mu, \phi, M, \rho_0)\) be a linear XEB scheme on \(n\) qubits. Suppose that the twirling map corresponding to \(\mu\) is a \(\gamma\)-approximate twirl with respect to the \(\|\cdot\|_2\) norm, and that \(\phi\) maps to probabilistic mixtures of unitaries:

\[
\phi(g) = \int d\nu_g u
\]

for probabilistic measures \(\nu_g\) on \(SU(d)\). Additionally, suppose that the implementation map satisfies

\[
E_{g \sim \mu} E_{u \sim \nu_g} \|\phi(g) - u\|_\diamond \leq \delta.
\]

(3.17)

If \(\delta \leq \frac{1-\gamma}{\pi^2}\), then there exists \(A, B \in \mathbb{R}\) and \(p \in [1-2\delta, 1]\), such that

\[
|q_R(m) - (A + Bp^m)| \leq 16 \times 2^{\frac{5}{2}n}(\gamma + 6\delta)^m.
\]

(3.18)

Compared to Theorem 5, Theorem 6 eliminates the exponential prefactor \(2^{n/2}\) on the noise level tolerance, with a slightly tighter restriction on the definition of the noise level \(\delta\) itself. However, \(\delta\) can be well approximated under an error model that can be readily expressed as mapping to mixtures of unitaries. We present the proof of Theorem 6 in Appendix D.1.

In general, it is much easier to prove a spectral gap than a gap with respect to the induced diamond norm. We have the following.

Definition 7. Let \(\mu\) be a probabilistic measure on a finite group \(G\), and \(A := \text{Supp}(\mu) \subseteq G\) be its support. We say that \(\mu\) is strongly scrambling if \(A\) is not contained in any (left or right) coset of a proper subgroup of \(G\), or equivalently if \(A^{-1}A\) generates \(G\).

Theorem 8. Assume that \(\mu\) is strongly scrambling over the \(n\)-qubit Clifford group. Then,

\[
\|\Lambda(\mu) - \Lambda(\mu_H)\|_2 < 1,
\]

where \(\mu_H\) is Haar measure on \(SU(2^n)\).

The proof of Theorem 8 can be found in Appendix D.2. That the 1-D chain and 2-D grid Clifford XEB schemes satisfy Definition 7 is straightforward and left as an exercise for the reader.

4 Discussion

In this work we propose Clifford XEB, a benchmarking scheme for large-scale quantum devices using Clifford circuits. Clifford XEB is both efficient in quantum circuit size and necessary classical processing, and it has great promise for benchmarking quantum devices much larger than heretofore possible. We numerically show that this scheme is feasible for more than a thousand qubits and theoretically prove that the scheme for the Clifford circuits we propose yields a single exponential decay for sufficiently small errors.

It is worthwhile to discuss in more detail why we need to holistically benchmark a quantum processor with tens or hundreds of qubits in the first place. One may think that it is sufficient to only perform single- or two-qubit benchmarking to characterize the native gate set. Even some effects of cross-talk can be captured via simultaneous single- or two-qubit RB [26]. The authors of [4] study this topic in depth by performing Clifford RB on a 3-qubit processor. Interestingly they find that the 3-qubit RB result varies depending on the calibration procedure in a way that is not captured by simultaneous single- or two-qubit RB. This suggests there is additional information to be gained from
running multiqubit benchmarks. This is especially relevant as the number of qubits on processors continues to increase and the calibration process becomes increasingly convoluted. Furthermore, there are a variety of obstacles to overcome to be able to scale quantum processors to achieve fault-tolerance, not the least of which the limited size of dilution fridges, and any workaround could introduce errors that can only be detected by a multiqubit benchmark. Lastly, using a multiqubit benchmark helps verify that the quantum processor does not have large correlated errors [1] which are detrimental to quantum error correction.

There have been other proposals for holistically benchmarking quantum computers. The quantum volume experiment [27] compares different processors in terms of their overall computational capability rather than just gate quality, but it is not scalable with respect to the effective number of qubits. Many proposals for estimating the global Pauli errors [28, 29, 30, 31] provides more detailed characterizations of the error, but may be more suitable for detailed debugging rather than a holistic figure of merit. Additionally, there have been variants of RB for holistically benchmarking quantum computers in a scalable way, such as matchgate benchmarking [32], direct randomized benchmarking [33], and mirror randomized benchmarking [34]. However, these schemes have not been as widely adopted in the experimental community as conventional linear XEB due to the latter’s ease of hardware implementation. Our scheme boasts the exact same convenience but is also manifestly scalable. Another distinguishing feature is that we can prove Clifford XEB yields an exponential decay under a general error model. There is also constructions of exact unitary 2-designs using almost linear sized circuits (up to log factors) [35]. Specifically, they require $O(n \log n \log \log n)$ size circuits. In contrast, the Clifford circuits we consider in our work are exactly linear size [12]. In fact, these log factors can still be quite large in practice (For example, if $n = 50$, $\log_2 n \log_2 \log_2 n \approx 14$, a non-trivial overhead on the required error rates.). Finally, in [7], there is a mention of conducting linear XEB with Clifford circuits. However, they do not pursue this line of thought nor further elaborate on its theory. Here we study Clifford XEB in depth from both a numerical and theoretical standpoint.

Similar to mirror RB [34], instead of classical simulation, we could sequentially reverse the Clifford layers on the quantum circuit and observe the exponential decay of the population of the initial state with respect to the number of cycles. The proof of exponential decay in [22] can be readily applied to this scheme as well. Compared to the Clifford XEB, this mirrored variant does not need to go through a scrambling phase and would probably need a smaller number of cycles. However, each cycle now consists of twice the number of gates compared to the Clifford XEB, and so the error tolerance of the protocol is halved compared to Clifford XEB. We leave a more detailed comparison for future work. Note that since it shares the same ideal twirling map as Clifford XEB, it could also potentially suffer from slow scrambling, making large gate error rates unextractable.

We give some directions for future research. Although numerical experiments in Section 2 show that Clifford XEB can be used to benchmark large quantum processors up to a thousand qubits, ideally we want to implement Clifford XEB on an actual quantum device. Although there are currently few quantum processors with on the order of a hundred qubits, it would already be interesting to test Clifford XEB on real quantum devices with tens of qubits. Given the auspicious results from numerical simulations, we expect that Clifford XEB can benchmark quantum processors with more qubits than is promised by theory.

One important practical question is to build a benchmarking workflow based on Clifford XEB. Since it benchmarks circuits of linear size, the Clifford XEB can serve as an overall figure of merit of the entire processor but also serve as a guideline for subsequent, more detailed benchmarking schemes.
Possible subsequent experiments include cycle benchmarking [28] or individual RB experiments. One can even apply Clifford XEB to subregions on a quantum chip as a measure of heterogeneous performance. We leave to future work to build such a practical workflow to efficiently obtain useful information about a quantum processor.

It would also be interesting to investigate, both from a theoretical and an experimental perspective, the mixing rate given a Clifford XEB scheme. Estimating the mixing rate is an essential step since it serves as a threshold below which the decay rate measured is not necessarily related to the gate error. For the numerical simulation on the $35 \times 35$ grid, it took half a day to gather the noisy linear XEB results, but two days and a half to get the mixing rate. This is because the fluctuation of the ideal linear XEB value is purely from the randomness of the circuits chosen and is much larger than that of the noisy experiments. A faster algorithm that can extract the decay exponent without having to sample millions of random circuits would be needed to unblock this bottleneck. Towards this end, there have been studies on the spectral cap from random Clifford circuits from a Markov chain perspective [36, 37, 14, 38, 39]. We leave it to future work to incorporate such methods to better estimate the spectral gap for particular random Clifford circuit ensembles.

Of course, it would also be useful to improve the constants in the number of qubits we can benchmark given by Equation (3.12). It would also be useful to realize a $\gamma$-approximate twirl in a way that is more natural for topologies that are more relevant for hardware, such as the 2-D grid. We could also consider topologies tailored for realizing quantum error correction, such as the honeycomb lattice [40]. Furthermore, we would love to see more work on how to make minimal assumptions on the gate errors to obtain a theoretical guarantee for a much higher number of qubits. For instance, most gates in superconducting circuits are limited by decoherence, which implies incoherent errors are dominant [23].

A possible alternative scheme to Clifford XEB is to use other gate sets that are classically simulable. If we can construct $\gamma$-approximate twirls using other gate sets, e.g., matchgates [41], then we can formulate XEB schemes that might be more relevant in other experimental settings.

Acknowledgements

We would like to thank Yaoyun Shi for helpful discussions and comments. DD would like to thank God for all of His provisions.

A Clifford Approximate Twirl Construction

For completeness, we describe explicitly the sequence of gates to construct the Clifford approximate twirl given in [12, Figure 1].

A.1 Description

We first define the following procedures. Suppose we have $n$ qubits in a star topology, where we denote the center qubit as qubit 1.

- **Perform a $\mathcal{P}_n$ twirl**: Apply a random Pauli gate on each qubit.

- **Perform a $\mathcal{C}_1/\mathcal{P}_1$ twirl**: Apply a gate uniformly chosen from the set \{I, S, H, SH, (SH)², SHS\}, with S being the single-qubit phase gate and H the Hadamard gate. Alternatively this could
be implemented by applying $(SH)^j$ with $j$ uniformly chosen from $\{0, 1, 2\}$, according to [12].

- **Conjugate qubit 1 by a random XOR:** For each qubit $2, \cdots, n$, apply a CNOT gate between this qubit and qubit 1 with probability $3/4$.

We now describe how to implement the Clifford approximate twirl:

1. Perform a $\mathcal{P}_n$ twirl.
2. Perform $\mathcal{C}_1/\mathcal{P}_1$ twirl on all of the qubits.
3. Conjugate qubit 1 by a random XOR.
4. Apply $H$ to qubit 1 and $\mathcal{C}_1/\mathcal{P}_1$ twirl the other qubits.
5. Conjugate qubit 1 by a random XOR.
6. Apply $H$ to qubit 1 and $\mathcal{C}_1/\mathcal{P}_1$ twirl the other qubits.
7. Apply $S$ to the first qubit with probability $1/2$.
8. Conjugate qubit 1 by a random XOR.
9. $\mathcal{C}_1/\mathcal{P}_1$ twirl the first qubit.
10. To obtain a $\gamma$-approximate twirl, repeat steps 2 to 9 for $O(\log(1/\gamma))$ times.

A.2 Different Topologies

The only two-qubit gate operation involved in this construction is conjugating qubit 1 by a random XOR. This requires a star topology circuit. Furthermore, a direct implementation of this procedure requires a circuit depth of $\Omega(n)$. It is possible to implement it in depth $O(\log n)$ on a fully connected graph, as was shown in [12, Figure 3], but realistic quantum processors have limited connectivity. To remedy this, we describe how to implement the random XOR circuit on a general connectivity graph. As a side result we also show that we can achieve depth $O(\log n)$ on a binary tree connectivity graph.

It is easy to see that in the computational basis, the conjugation by XOR operation effectively applies an X-gate to qubit 1 conditioned on the XOR of the random set. We therefore give a construction on a general graph by finding a spanning tree and iteratively computing the XOR of the random set through each level of the tree. Given the graph is connected, we can choose an arbitrary qubit to be the root (qubit 1), and form a spanning tree rooted from qubit 1. Each qubit $v$ apart from qubit 1 is then assigned a depth $d(v)$ as its distance from qubit 1 on the spanning tree, and a degree $\delta(v)$. Moreover, denote $D = \max_v d(v)$ and $\Delta = \max_v \delta(v)$ be the depth and degree of the spanning tree. Now, for each of the qubits other than qubit 1, add it to a set $S$ with probability $3/4$. We call a qubit $v$ active, if at least one of the qubits in the subtree rooted at $v$ (including $v$) is in $S$. We then perform the following operations:

1. Do the following sequentially for each $l$ from $D - 1$ to 1:
   
   (a) Do the following in parallel for each active qubit $x$ at depth $l$:
i. If \( x \) is not in \( S \), it must have an active child \( y \). Choose an arbitrary such \( y \) and apply a CNOT gate from \( x \) to \( y \).

ii. Apply CNOT gates from each active child of \( x \) to \( x \) sequentially.

2. Apply CNOT gates from all active qubits at depth 1 to the root sequentially.

3. Apply the CNOT gates generated in step 1 again but in reverse order.

To see why this circuit works, we first assume that every qubit is in the computational basis before the circuit, and the general case follows from linearity. For step 1, one can see by induction that all qubits of depth \( l' \) will have a value that equals to the XOR of qubits in its subtree that are contained in \( S \) when the loop with \( l = l' \) finishes. Step 1(a)i ensures that if \( x \) is not in \( S \), its value will be canceled after step 1(a)ii. Step 2 ensures that qubit 1 gets the corresponding XOR values from its children, and step 3 “uncomputes” the intermediate results in other qubits.

The total depth of the circuit can be upper bounded by \( 2 - D \Delta \). To see this, observe that step 3 takes the same depth as step 1, which takes \((D - 1) \cdot \Delta \) layers for sequentially executing \( D - 1 \) repetitions of step 1(a), each takes at most \( \Delta \) layers. Therefore, the depth of the circuit greatly depends on the spanning tree, which in turn depends on the underlying connectivity graph. When the spanning tree can be chosen to be a perfect binary tree, the random XOR can be done in \( \Theta(\log(n)) \) depth. For hardware relevant graphs such as the 1-d chain and the 2-d grid, the circuit depth is upper bounded by \( O(D) \) as the degree of the spanning tree is upper bounded by the degree of the connectivity graph. In this case, the depth of the spanning tree can be minimized to the radius of the connectivity graph \( r(G) = \min_u \max_v d(u, v) \) by rooting at \( \arg \min_u \max_v d(u, v) \) and taking the spanning tree to consist of the shortest paths from each node to the root. The radius of the 1-d chain and the 2d grid is \( \Theta(n) \) and \( \Theta(\sqrt{n}) \) respectively. One can see that this is asymptotically tight as it takes \( \Omega(r(G)) \) steps to propagate information from an arbitrary qubit to a fixed one.

## B  Expectation and Variance of linear XEB

It is observed in Section 2 that the Clifford XEB value converges to 1 and 0 for the ideal cases and the noisy cases respectively. Similar behavior has also been observed in linear XEB experiments [1, 2]. In this section we analyze the expectation and variance for ideal and noisy Clifford XEB values when the number of cycles is sufficiently large. We assume that for Clifford XEB and linear XEB, the first four moments of the distribution of the random gates converge to those of the uniform distribution over the Clifford group and the Haar measure over the unitary group respectively.

### B.1  Ideal Case

For any circuit \( C \in SU(2^n) \), we define

\[
\beta_C = 2^n \sum_x |\langle x|C|0^n \rangle|^4, \tag{B.1}
\]

which is the ideal linear XEB value plus 1. When \( C \) is Haar random in \( SU(2^n) \), \( C \langle 0^n \rangle \) is a uniformly random unit vector in \( \mathbb{C}^{2^n} \). This vector can be parameterized as

\[
(e^{i\phi_1} \cos \theta_1, e^{i\phi_2} \sin \theta_1 \cos \theta_2, \ldots, e^{i\phi_D} \sin \theta_1 \sin \theta_2 \ldots \sin \theta_{D-1}) \tag{B.2}
\]
in the computational basis, where \( D = 2^n \) is the dimension. To compute the variance we claim we only need to consider the dependence on \( \theta_1 \) and \( \theta_2 \). The metric of the integral will contain two factors: the first factor is related to the spherical coordinates \( \theta \) and is proportional to \( \sin^{D-2} \theta_1 \sin^{D-3} \theta_2 \), and the second factor is related to the phase factors and is equal to

\[
\cos \theta_1 \times \sin \theta_1 \cos \theta_2 \times \sin \theta_1 \sin \theta_2 \cos \theta_3 \times \ldots \times \cos \theta_1 \cos \theta_2 \sin^{D-1} \theta_1 \sin^{D-2} \theta_2. \tag{B.3}
\]

Let \( t_1 = \cos^2 \theta_1, t_2 = \sin^2 \theta_1 \cos^2 \theta_2 \), and the distribution of \( t_1 \) and \( t_2 \) is given by

\[
\Pr[t_1, t_2] \propto \int d\theta_1 d\theta_2 \delta(t_1 - \cos^2 \theta_1) \delta(t_2 - \sin^2 \theta_1 \cos^2 \theta_2) \sin^{D-2} \theta_1 \sin^{D-3} \theta_2 \times \cos \theta_1 \cos \theta_2 \sin^{D-1} \theta_1 \sin^{D-2} \theta_2 \times \alpha(1 - t_1 - t_2)^{D-3}. \tag{B.4}
\]

for \( 0 \leq t_1, t_2 \leq 1, t_1 + t_2 \leq 1 \). One can normalize the distribution and have

\[
\Pr[t_1, t_2] = (D - 1)(D - 2)(1 - t_1 - t_2)^{D-3}. \tag{B.5}
\]

If we integrate over \( t_2 \), the distribution of \( t_1 \) is

\[
\Pr[t_1] = (D - 1)(1 - t_1)^{D-2}. \tag{B.6}
\]

By symmetry, for any \( x \neq y \) the value of \( |\langle x | C | 0^n \rangle|^2 \) and \( |\langle y | C | 0^n \rangle|^2 \) will satisfy this distribution by identifying them with \( t_1 \) and \( t_2 \) respectively. Then one can calculate the mean and variance of \( \beta_C \) as follows.

\[
\mathbb{E}_{C;\beta_C} = D \sum_x |\langle x | C | 0^n \rangle|^4 \\
= D \int_0^1 dt_1 \Pr[t_1] t_1^2 \\
= \frac{2D}{D + 1}, \tag{B.7}
\]

\[
\mathbb{E}_{C;\beta_C^2} = D^2 \mathbb{E}_{C} \left( \sum_x |\langle x | C | 0^n \rangle|^4 \right)^2 \\
= D^2 \mathbb{E}_{C} \sum_{x \neq y} |\langle x | C | 0^n \rangle|^4 |\langle y | C | 0^n \rangle|^4 + D^2 \mathbb{E}_{C} \sum_x |\langle x | C | 0^n \rangle|^8 \\
= D^3(D - 1) \int_0^1 dt_1 \int_0^{1-t_1} dt_2 \Pr[t_1, t_2] t_1^2 t_2^2 + D^3 \int_0^1 dt_1 \Pr[t_1] t_1^4 \\
= \frac{4D^2(D + 5)}{(D + 1)(D + 2)(D + 3)}, \tag{B.8}
\]

and so the variance is given by

\[
\mathbb{E}_{C;\beta_C^2} - (\mathbb{E}_{C;\beta_C})^2 = \frac{4D^2(D - 1)}{(D + 1)^2(D + 2)(D + 3)} = \frac{4}{D} + O(D^{-2}). \tag{B.9}
\]

From Equation (B.7) one can see that the ideal linear XEB value converges to \( \frac{D - 1}{D + 1} \), which is close to 1 for large \( D \). This also applies to Clifford XEB since the uniform distribution on the Clifford group forms a unitary 2-design.
When $C$ is a uniformly random Clifford gate, $C|0^n\rangle$ will be a uniformly random stabilizer state. We introduce the concept of $k$-neighbor of a stabilizer state $|\psi\rangle$, which is the set of stabilizer states $|\phi\rangle$ such that $|\langle\phi|\psi\rangle| = 2^{-k/2}$. It is known that every $n$-qubit stabilizer state has the same number of $k$-neighbors [42], which is denoted by $L_n(k)$. We know that for any stabilizer state $C|0^n\rangle$, the inner product $|\langle x|C|0^n\rangle|$ is $2^{-k/2}$ for $2^k$ choices of $x$, where $k$ is an integer, and for the remaining values of $x$ the inner product is 0. It is easy to see that $\beta_C = 2^{n-k}$ in this case.

For each $|x\rangle$, there are $L_n(k)$ $k$-neighbors, and each such state is in turn a $k$-neighbor of $2^k$ computational states. This means that there are $2^{n-k}L_n(k)$ stabilizer states that leads to $\beta_C = 2^{n-k}$. Let $N(n)$ be the total number of $n$-qubit stabilizer states. Then

$$\mathbb{E}_C \beta_C = \sum_k 2^{n-k} \frac{L_n(k)}{N(n)} = \sum_t 4^t \frac{L_n(n-t)}{N(n)}, \quad (B.10)$$

$$\mathbb{E}_C \beta_C^2 = \sum_k 4^{n-k} \frac{L_n(k)}{N(n)} = \sum_t 8^t \frac{L_n(n-t)}{N(n)}. \quad (B.11)$$

It is known [42, Thm. 16] that in the large $n$ limit,

$$c_1 2^{-t(t+5)/2} \leq \frac{L_n(n-t)}{N(n)} \leq c_2 2^{-t(t+3)/2} \quad (B.12)$$

for some constants $c_1$ and $c_2$. So $\mathbb{E}_C \beta_C$ and $\mathbb{E}_C \beta_C^2$ are both $\Theta(1)$. This means that the variance of $\beta_C$ is $\Theta(1)$ when $C$ is a uniformly random Clifford circuit.

### B.2 Noisy Case

In this section, we investigate the single exponential decay behavior in the noisy case. In particular, we give estimates for the coefficients $A$ and $B$ in Theorem 4, showing that $A = 0$ and $B = O(1)$.

Following the universal randomized benchmarking (URB) framework [22], we can write the Clifford XEB experiment results as

$$q_R(m) = -1 + 2^n \text{tr}[|0^n\rangle\langle 0^n| (\Lambda_R^m(D_M)(\rho_0))],$$

where the noisy twirling map $\Lambda_R$ is defined as a map on quantum channels:

$$\Lambda_R : \mathcal{C} \mapsto \mathbb{E}_{g \sim \text{Haar}}[g^\dagger \circ \mathcal{C} \circ \phi(g)].$$

Now, using the notations in the proof of [22, Theorem 8], $\Lambda_R$ is regarded as a perturbed version of $\Lambda_R^* \Lambda_R$ and has a gapped spectrum leading to the single exponential decay in an appropriate parameter range. Specifically, define

$$X_1 = \frac{\langle \cdot , D\rangle_{SO} D}{\langle D,D\rangle_{SO}} + \frac{\langle \cdot , I-D\rangle_{SO} (I-D)}{\langle I-D, I-D\rangle_{SO}} (I-D), \quad X_2 = I - X_1,$$

One can verify that $X_i^2 = X_i$ for $i = 1, 2$ and

$$X_i \Lambda_R^* X_j = 0$$

when $i \neq j$. The matrix perturbation result states that when $||\Lambda_R^* - \Lambda_{\text{Haar}}||_0 < \gamma$, $||\Lambda_R^* - \Lambda_R||_0 \leq \delta$ and $\delta < \frac{1-\gamma}{11}$, there exists twirling maps $L, R$ such that
\[ L = R^{-1}, \quad |||L|||_\diamond \leq 4, \quad |||R|||_\diamond \leq 4; \]
\[ X_i R \Lambda R L X_j = 0 \text{ when } i \neq j; \]
\[ |||X_2 R \Lambda R L X_2|||_\diamond \leq \gamma + 6 \delta; \]
\[ |||X_1 R \Lambda R L X_1 - X_1|||_\diamond \leq 2 \delta. \]

Let \( A'_i = X_i R \Lambda R L X_i \). It can be further proven that \( A'_1 \) has one eigenvalue 1 and another real eigenvalue lying in \([1 - 2 \delta, 1]\). Then
\[ \Lambda^m_R = L(A'_1)^m R + L(A'_2)^m R \]
and
\[ q_R(m) = -1 + 2^n \text{tr}[[0^n]\langle 0^n| (\Lambda^m_R(\mathcal{D}_M) (\rho_0))] \]
\[ = -1 + 2^n \left[ \text{tr}[\langle 0^n| (L(A'_1)^m R(\mathcal{D}_M)(\rho_0)) + \text{tr}[\langle 0^n| (L(A'_2)^m R(\mathcal{D}_M)(\rho_0))\right]. \quad (B.13) \]

Since
\[ |||A'_2|||_\diamond \leq \gamma + 6 \delta, \]
the second term vanishes quickly, and the single exponential decay is given by the first term.

We assumed \( p < 1 \), so \( A \) is well-defined and is the contribution corresponding to the eigen-channel of \( \Lambda_R \) with eigenvalue 1. It is easy to see that this eigen-channel is the depolarization channel \( \mathcal{D}(\rho) = \frac{I}{D} \):
\[ \Lambda_R(\mathcal{D})(\rho) = \int d\mu \omega^\dagger(g) \circ \mathcal{D}(\phi(g)(\rho)) = \int d\mu \omega^\dagger(g)(\frac{I}{D}) = \frac{I}{D} \]

since the inverting maps \( \omega^\dagger(g) \) are all unital. We then have
\[ A = 2^n \text{tr}[[0^n|\langle 0^n|L(\mathcal{D}(\rho)] - 1 = 0. \]

We thereby obtain the first conclusion \( A = 0 \). Note that this holds in general when \( \Lambda_R \) only has only one eigenvalue with modulus 1: This ensures that \( \Lambda^m_R(\mathcal{D}_M) \rightarrow \mathcal{D} \) when \( m \rightarrow \infty \). This criterion does not pose any requirement on the state preparation or the measurement.

Both \( \Lambda_R \) and \( \Lambda_R^* \) are twirling maps; they map channels to channels. Consequently, they map differences of channels to differences of channels. It is thus more convenient to restrict them on the space spanned by the differences of channels. In this case one can define \( X'_1 \) to be
\[ X'_1 = \frac{\langle \cdot, \mathcal{I} - \mathcal{D} \rangle_{SO}}{\langle \mathcal{I} - \mathcal{D}, \mathcal{I} - \mathcal{D} \rangle_{SO}} (\mathcal{I} - \mathcal{D}). \]

The perturbation result tells us that
\[ |B| = |2^n \text{tr}[[0^n|\langle 0^n|L \cdot X'_1 \cdot R(\mathcal{D}_M)(\rho)|] | \]
\[ \leq 2^n \cdot |||L \cdot X'_1(R(\mathcal{D}_M))|||_\diamond \]
\[ \leq 2^n |||(R(\mathcal{D}_M), \mathcal{I} - \mathcal{D})_{SO} \cdot L(\mathcal{I} - \mathcal{D}) \|_{\langle \mathcal{I} - \mathcal{D}, \mathcal{I} - \mathcal{D} \rangle_{SO}} \]
\[ = \frac{2^n}{\langle \mathcal{I} - \mathcal{D}, \mathcal{I} - \mathcal{D} \rangle_{SO}} |||L(\mathcal{I} - \mathcal{D})|||_\diamond |\langle \mathcal{D}_M, R(\mathcal{I} - \mathcal{D}) \rangle_{SO}|. \]
For sake of simplicity we denote $\mathcal{N} := R^l(\mathcal{I} - \mathcal{D})$. We now bound the quantity $|\langle \mathcal{D}_M, \mathcal{N} \rangle_{SO}|$.

Let $\mathcal{P}: \rho \rightarrow \sum_i \langle i | \rho | i \rangle | i \rangle \langle i |$ be the dephasing channel. One can check that it is self-adjoint with respect to the super operator inner product. Therefore

$$\langle \mathcal{D}_M, \mathcal{N} \rangle_{SO} = \langle \mathcal{P} \circ \mathcal{D}_M, \mathcal{N} \rangle_{SO} = \langle \mathcal{D}_M, \mathcal{P} \circ \mathcal{N} \rangle_{SO}.$$  

As $\|\mathcal{P} \circ \mathcal{N}\|_\circ \leq \|\mathcal{N}\|_\circ$ due to data processing inequality, we can assume without generality that $\mathcal{N}$ is a quantum-classical superoperator: that is, there exists a set of operators $\{N_i\}_i$ such that

$$\mathcal{N}: \rho \rightarrow \sum_i \text{tr}[N_i \rho] | i \rangle \langle i |.$$

Then

$$\langle \mathcal{D}_M, \mathcal{N} \rangle_{SO} = \sum_{P_i \in \mathcal{P}^\otimes n} \text{tr}[\mathcal{D}(P_i) N(P_i)]$$

$$= \sum_{P_i \in \mathcal{P}^\otimes n} \sum_{j \in \{0,1\}^n} \text{tr}[M_j P_i] [N_j P_i]$$

$$= 2^n \sum_j \text{tr}[M_j N_j]$$

$$= 4^n \sum_j \frac{\text{tr}[M_j]}{2^n} \cdot \text{tr}[M_j N_j],$$

where $\tilde{M}_i = \frac{M_i}{\text{tr}[M_i]}$. Since $\{M_i\}_i$ is a POVM, we have $M_i \geq 0$ and $\sum_i M_i = I$, and therefore $\tilde{M}_i$ is a quantum state. We argue that $|\langle \mathcal{D}_M, \mathcal{N} \rangle_{SO}| \leq 4^n \|\mathcal{N}\|_\circ$. Assume otherwise, then there exists $i$ such that

$$|\text{tr}[\tilde{M}_i N_i]| > \|\mathcal{N}\|_\circ.$$

However this is not possible as

$$\|\mathcal{N}\|_\circ \geq |\mathcal{N}(\tilde{M}_i)|_{tr} = \sum_j |\text{tr}[\tilde{M}_i N_j]| \geq |\text{tr}[\tilde{M}_i N_i]|.$$

We then calculate $\langle \mathcal{I} - \mathcal{D}, \mathcal{I} - \mathcal{D} \rangle_{SO}$. By linearity we have

$$\langle \mathcal{I} - \mathcal{D}, \mathcal{I} - \mathcal{D} \rangle_{SO} = \langle \mathcal{I}, \mathcal{I} \rangle_{SO} - 2\langle \mathcal{D}, \mathcal{I} \rangle_{SO} + \langle \mathcal{D}, \mathcal{D} \rangle_{SO}.$$  

For any quantum channel $\mathcal{C}$,

$$\langle \mathcal{D}, \mathcal{C} \rangle_{SO} = \sum_{P_i \in \mathcal{P}^\otimes n} \text{tr}[\mathcal{D}(P_i) \mathcal{C}(P_i)]$$

$$= \sum_{P_i \in \mathcal{P}^\otimes n} \text{tr}[2^{-n} \cdot I \text{tr}[P_i] \mathcal{C}(P_i)]$$

$$= \text{tr}[2^{-n} \cdot 2^n \cdot \mathcal{C}(I)] = 2^n.$$  

On the other hand,

$$\langle \mathcal{I}, \mathcal{I} \rangle_{SO} = \sum_{P_i \in \mathcal{P}^\otimes n} \text{tr}[\mathcal{I}(P_i) \mathcal{I}(P_i)]$$

$$= \sum_{P_i \in \mathcal{P}^\otimes n} \text{tr}[I] = 4^n \cdot 2^n = 8^n.$$
Therefore

\[ (I - D, I - D)_{SO} = 8^n - 2^n. \]

Plugging everything in we have

\[ B \leq \|N\|_\diamond \cdot \|L(I - D)\|_\diamond \frac{4^n}{4^n - 1}. \]

Since \( N = R^\dagger(I - D) \), we have

\[ |B| \leq ||L||_\diamond \cdot ||R||_\diamond \cdot \|I - D\|_\diamond \frac{4^n}{4^n - 1} \leq 16 \cdot 4 \cdot 2 \leq 128, \]

proving that \( B = O(1) \).

**C Circuit Lower Bounds for \( \gamma \)-Approximate Twirl**

In this section we prove lower bounds on the circuit required to implement \( \gamma \)-approximate twirls. We state our result in the form of a theorem:

**Theorem 9.** Suppose that the twirling map corresponding to a measure \( \mu \) on \( SU(2^n) \) is a \( \gamma \)-approximate twirl. Suppose we implement unitaries \( g \in SU(2^n) \) sampled from \( \mu \) by compiling them using single- and two-qubit gates, and let \( s(g) \) and \( t(g) \) be the number of gates and the depth for the circuit that implements \( g \). Then

\[ \mathbb{E}_{g \sim \mu} s(g) = \Omega((1 - \gamma)n), \mathbb{E}_{g \sim \mu} t(g) = \Omega((1 - \gamma) \log n). \quad (C.1) \]

**Proof.** This proof is adapted from [35]. We set the channel \( N \) so that \( N(\rho) = Z_1 \rho Z_1 \) where \( Z_1 \) is the Pauli \( Z \) gate applied to the first qubit. Then we know that the twirling map \( \Lambda_{\text{Haar}} \) maps \( N \) to the following channel:

\[ \Lambda_{\text{Haar}}(N) : \rho \mapsto \frac{1}{4^n - 1} \sum_{\sigma_p \neq 0^n} \sigma_p \rho \sigma_p = \frac{4^n}{4^n - 1} \times \frac{I}{2^n \text{tr}[\rho]} - \frac{1}{4^n - 1} \rho. \quad (C.2) \]

Let \( D \) be the completely depolarizing channel on \( n \) qubits

\[ D(\rho) = \frac{I}{2^n \text{tr}[\rho]}, \quad (C.3) \]

then one can see that

\[ \|D - \Lambda_{\text{Haar}}(N)\|_\diamond = O(4^{-n}). \quad (C.4) \]

By definition

\[ \Lambda_{\mu}(N) : \rho \mapsto \mathbb{E}_{g \sim \mu} g^\dagger Z_1 g \rho g^\dagger Z_1 g. \quad (C.5) \]

Let \( f(g) \) be the size of the support of \( g^\dagger Z_1 g \), i.e., the number of qubits that \( g^\dagger Z_1 g \) acts nontrivially on. For any \( 1 \leq j \leq n \), define \( p_j \) as the probability that qubit \( j \) is in the support of \( g^\dagger Z_1 g \) for \( g \)
sampled according to $\mu$. Assuming $p_j \leq 1/2$, we have
\[
\gamma \geq \| \Lambda_\mu(N) - \Lambda_{\text{Haar}}(N) \|_\circ \\
\geq \| \Lambda_\mu(N) - D \|_\circ - O(4^{-n}) \\
\geq \| (\Lambda_\mu(N) - D)(\langle 0^n | \langle 0^n \|_1 - O(4^{-n}) \\
\geq \| \text{tr}_j[(\Lambda_\mu(N) - D)(\langle 0^n | \langle 0^n \|_1 - O(4^{-n}) \\
= \| p_j \rho_j + (1 - p_j) |0\rangle \langle 0| - I/2 \|_1 - O(4^{-n}) \\
\geq \| p_j |1\rangle \langle 1| + (1 - p_j) |0\rangle \langle 0| - I/2 \|_1 - O(4^{-n}) \\
= 1 - 2p_j - O(4^{-n}),
\] (C.6)
where $\text{tr}_j[\cdot]$ means tracing over all qubits other than qubit $j$, and in Equation (C.6) $\rho_j$ refers to the state of qubit $j$ conditioned that some unitary is applied to $j$ in channel $\Lambda_\mu(N)$. This implies
\[
p_j \geq \frac{1 - \gamma}{2} - O(4^{-n}).
\] (C.8)
Obviously Equation (C.8) is also satisfied when $p_j > 1/2$, so it always holds. Then
\[
\mathbb{E}_{g \sim \mu} f(g) = \sum_j p_j \geq \frac{1 - \gamma}{2} n - O(n4^{-n}).
\] (C.9)
Note that for any $g$,
\[
s(g) \geq f(g) - 1, t(g) \geq \log f(g) \geq \frac{f(g) - 1}{n - 1} \log n,
\] (C.10)
so we have
\[
\mathbb{E}_{g \sim \mu} s(g) = \Omega((1 - \gamma)n), \mathbb{E}_{g \sim \mu} t(g) = \Omega((1 - \gamma) \log n).
\] (C.11)

\section*{D Spectral Norm Proofs}

\subsection*{D.1 Proof of Theorem 6}
In this section we give a proof of Theorem 6. Specifically, following the proof of Theorem 16 in [22], it suffices to prove the following.

\textbf{Proposition 10.} Let $\Lambda^* = \int d\mu \omega^\dagger(g) \circ \cdot \omega(g)$ be the ideal twirling map associated to the probabilistic measure $\mu$, and $\Lambda = \int d\mu \omega^\dagger(g) \circ \cdot \phi(g)$ be its noisy implementation. Moreover, suppose that $\phi$ maps to mixtures of unitaries, that is, for each $g$ there exists a probabilistic measure $\nu_g$ on $SU(d)$ such that
\[
\phi(g) = \int d\nu_g u.
\]
Then
\[
\|\|\Lambda^* - \Lambda\|\|_2 \leq \mathbb{E}_{g \sim \mu} \mathbb{E}_{u \sim \nu_g} \| \omega(g) - u \|_\circ.
\]
Proof. It can be verified that
\[ |||\Lambda^* - \Lambda|||_2 \leq \int d\mu ||\omega(g) - \phi(g)||_2 \leq \int d\mu \int d\nu_g ||\omega(g) - u||_2.\]

It then suffices to prove that \( ||\omega(g) - u||_2 \leq ||\omega(g) - u||_{tr} \) for unitary channels \( \omega(g) \) and \( u \) since it is known that \( ||\omega(g) - u||_{tr} \leq ||\omega(g) - u||_o.\)

Assume without loss of generality that \( \omega(g) = \text{Id} \) and \( u = U \cdot U^\dagger.\) Assume further that \( U \) is diagonalized over the computational basis: \( U = \sum_j e^{i\lambda_j} |j\rangle \langle j|.\) Then the trace norm and the 2-norm can then be computed explicitly as
\[ ||\text{Id} - u||_{tr} = ||\text{Id} - u||_2 = 2 \max_{i,j} \sin \frac{\lambda_i - \lambda_j}{2}.\]

\[ \square \]

D.2 Spectrally Gapped Schemes

In this section we provide a proof to Theorem 8. Our main tool is random walks on finite groups [43]. We first make the following definition:

Definition 11. Let \( G \) be a finite group. We say a measure \( \mu \) over \( G \) is scrambling over \( G \) if \( A := \text{Supp}(\mu) \) generates \( G \), and \( A \) is not contained in a coset of a proper normal subgroup of \( G \).

This is interesting because from [43], we can show that this measure converges to the uniform distribution under iteration, where iteration is defined as
\[ (\mu_1 \ast \mu_2)(g) := \int d\mu_2 \mu_1(\mu_2 g g^{-1}) \mu_2(g_2). \] (D.1)

We state this fact as a proposition:

Proposition 12 (Proposition 2.3 and Theorem 2.1 of [43]). Let \( G \) be a finite group and let \( \mu \) be a measure that is scrambling over \( G \). Then, \( \mu \) converges to the uniform distribution \( \mu_U \) over \( G \) under iteration.

We can now prove the following theorem.

Theorem 13. Let \( \mu \) be an inverse symmetric measure (that is, \( \mu(g) = \mu(g^{-1}) \) for all \( g \)) that is scrambling over the \( n \)-qubit Clifford group \( \mathcal{C}(n) \). Then,
\[ |||\Lambda(\mu) - \Lambda(\mu_H)|||_2 < 1, \] (D.2)

where \( \mu_H \) is Haar measure on \( SU(2^n) \).

Proof. Treating the twirling maps as linear maps on real superoperators, \( \Lambda(\mu_H) \) is a projector onto the space spanned by the identity channel \( \text{id} \) and the completely depolarizing channel \( \text{dep} \). These two channels are also eigenvectors of \( \Lambda(\mu) \) with eigenvalue 1. Since the singular values of \( \Lambda(\mu) \) are at most 1, \( |||\Lambda(\mu) - \Lambda(\mu_H)|||_2 \) is therefore the third largest singular value of \( \Lambda(\mu) \).

Now, by the argument in the proof of Theorem 8 in [22], since \( \mu \) is inverse symmetric, \( \Lambda(\mu) \) is self-adjoint with respect to the SO inner product on real superoperators. Hence, it is diagonalizable.
by the spectral theorem. This implies the third largest singular value of $\Lambda(\mu)$ is third largest element of the elementwise absolute value of $\text{spec}(\Lambda(\mu))$.

Next, it is easy to see that
\[ \Lambda(\mu^*k) = \Lambda(\mu)^k. \]  
(D.3)

Next, since $\Lambda(\mu)$ is diagonalizable, the eigenvalues of $\Lambda(\mu)^k$ are the $k$-th powers of the eigenvalues of $\Lambda(\mu)$. Two eigenvalues of $\Lambda(\mu)$ are 1, and every other eigenvalue has to be of magnitude at most 1.

Assume for contradiction that there is a third eigenvalue of $\Lambda(\mu)$ with magnitude 1. Then, for any $k$, $\Lambda(\mu)^k$ always has at least three eigenvalues with magnitude 1. However, by Proposition 12, for all $g \in C(n)$,
\[ \lim_{k \to \infty} \mu^k(g) - \mu_U(g) = 0. \]  
(D.4)

Thus, for arbitrarily large $k$, the difference is arbitrarily small. Furthermore,
\[ |||\Lambda(\mu^*k) - \Lambda(\mu_U)|||_2 \leq \max_{g \in C(n)} (\mu^k(g) - \mu_U(g))|C(n)|. \]  
(D.5)

Thus, for arbitrarily large $k$, the $|||\cdot|||_2$ norm is also arbitrarily small. Now, since groups are closed under inverses, $\mu_U$ is symmetric under inverse. Thus, $\Lambda(\mu_U)$ is self-adjoint. Furthermore, since $\Lambda(\mu)$ is self-adjoint, $\Lambda(\mu^*k) = \Lambda(\mu)^k$ is also self-adjoint. By Weyl’s inequality on matrix perturbation, the eigenvalues of $\Lambda(\mu^*k)$ and $\Lambda(\mu_U)$ can be paired up with a difference of at most $|||\Lambda(\mu^*k) - \Lambda(\mu_U)|||_2$. However, this can be arbitrarily small as $k \to \infty$, and since $C(n)$ is a unitary 2-design, two eigenvalues of $\Lambda(\mu_U) = \Lambda(\mu_H)$ are 1 and the rest are 0. This clearly contradicts $\Lambda(\mu)^k$ having at least three eigenvalues with magnitude 1. Hence we conclude $\Lambda(\mu)$ has only two eigenvalues with magnitude 1.

This means the third largest element of the elementwise absolute value of $\text{spec}(\Lambda(\mu))$ is strictly less than 1, which is the desired conclusion.

We can actually leverage this theorem to strongly scrambling measures $\mu$ that are not inverse symmetric. Define the inverse measure
\[ \mu^{-1}(g) := \mu(g^{-1}). \]  
(D.6)

We first show the following.

**Proposition 14.** Let $\mu$ be a probabilistic measure on a finite group $G$. Let $A := \text{Supp}(\mu)$. Then the following are equivalent:

1. $A$ does not lie in any (left or right) coset of a proper subgroup of $G$;
2. $\langle A^{-1}A \rangle = G$;
3. $\mu^{-1} * \mu$ is scrambling over $G$.

**Proof.** We follow the directions $1 \Leftrightarrow 2 \Leftrightarrow 3$.

- $1 \Rightarrow 2$. Assume otherwise that $\langle A^{-1}A \rangle = H < G$. Then for any $a, b \in A$, we have
  \[ a^{-1}b \in H \Rightarrow b \in aH \Rightarrow A \subseteq aH, \]
  contradicting with 1.
• $2 \Rightarrow 1$. Note that $aH = (aHa^{-1})a$, and a left cost of a subgroup is at the same time a right coset of a conjugate subgroup. Assume that $A \subseteq aH$ for some $a \in G, H < G$. Then $\langle A^{-1}A \rangle \subseteq H < G$, contradicting with $2$.

• $2 \Rightarrow 3$. It is easy to verify that $\text{Supp}(\mu^{-1} \ast \mu) = A^{-1}A$. It suffices to prove that $A^{-1}A$ does not lie in any coset $aH$ of a proper normal subgroup $H \triangleleft G$. Assume otherwise, then either $1 \in aH = H$ and $\langle A^{-1}A \rangle \subseteq H < G$ contradicting with $2$, or $1 \notin aH$ contradicting $1 \in A^{-1}A$.

• $3 \Rightarrow 2$. Follows by definition.

\[ \text{Proof of Theorem 8.} \] Since $\mu$ is strongly scrambling, then by Proposition 14 we know that $\mu^{-1} \ast \mu$ is scrambling. Without loss of generality, assume the former. By Theorem 13,

\[ \| | \Lambda(\mu^{-1} \ast \mu) - \Lambda(\mu_H) ||_2 < 1. \] (D.7)

That is, the third largest magnitude of the spectrum of $\Lambda(\mu^{-1} \ast \mu)$ is less than 1. Since $\Lambda(\mu^{-1} \ast \mu) = \Lambda(\mu)\Lambda(\mu)^\dagger$, we conclude the third largest singular value of $\Lambda(\mu)$ is less than 1, which is the desired conclusion. \hfill \square

References

[1] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[2] Yulin Wu, Wan-Su Bao, Sirui Cao, Fusheng Chen, Ming-Cheng Chen, Xiawei Chen, Tung-Hsun Chung, Hui Deng, Yajie Du, Daojin Fan, et al. Strong quantum computational advantage using a superconducting quantum processor. *Physical review letters*, 127(18):180501, 2021.

[3] Hugh Collins and Kortney Easterly. Ibm unveils breakthrough 127-qubit quantum processor. *IBM Newsroom*, Nov 2021.

[4] David C McKay, Sarah Sheldon, John A Smolin, Jerry M Chow, and Jay M Gambetta. Three-qubit randomized benchmarking. *Physical review letters*, 122(20):200502, 2019.

[5] Xun Gao, Marcin Kalinowski, Chi-Ning Chou, Mikhail D Lukin, Boaz Barak, and Soonwon Choi. Limitations of linear cross-entropy as a measure for quantum advantage. *arXiv preprint arXiv:2112.01657*, 2021.

[6] Xiao Mi, Pedram Roushan, Chris Quintana, Salvatore Mandrà, Jeffrey Marshall, Charles Neill, Frank Arute, Kunal Arya, Juan Atalaya, Ryan Babbush, et al. Information scrambling in quantum circuits. *Science*, 374(6574):1479–1483, 2021.

[7] Yunchao Liu, Matthew Otten, Roozbeh Bassirianjahromi, Liang Jiang, and Bill Fefferman. Benchmarking near-term quantum computers via random circuit sampling. *arXiv preprint arXiv:2105.05232*, 2021.

27
[8] Daniel Gottesman. *Stabilizer codes and quantum error correction*. PhD thesis, Caltech, 1997. eprint: quant-ph/9705052.

[9] Joseph Emerson, Robert Alicki, and Karol Życzkowski. Scalable noise estimation with random unitary operators. *Journal of Optics B: Quantum and Semiclassical Optics*, 7(10):S347, 2005.

[10] Benjamin Lévi, Cecilia C. López, Joseph Emerson, and D. G. Cory. Efficient error characterization in quantum information processing. *Phys. Rev. A*, 75:022314, Feb 2007.

[11] Jonas Helsen, Ingo Roth, Emilio Onorati, Albert H Werner, and Jens Eisert. A general framework for randomized benchmarking. *arXiv preprint arXiv:2010.07974*, 2020.

[12] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1):012304, 2009.

[13] Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, July 2021.

[14] Aram W Harrow and Richard A Low. Random quantum circuits are approximate 2-designs. *Communications in Mathematical Physics*, 291(1):257–302, 2009.

[15] Patrick Hayden and John Preskill. Black holes as mirrors: quantum information in random subsystems. *Journal of high energy physics*, 2007(09):120, 2007.

[16] Yasuhiro Sekino and Leonard Susskind. Fast scramblers. *Journal of High Energy Physics*, 2008(10):065, 2008.

[17] Nima Lashkari, Douglas Stanford, Matthew Hastings, Tobias Osborne, and Patrick Hayden. Towards the fast scrambling conjecture. *Journal of High Energy Physics*, 2013(4):1–33, 2013.

[18] Winton Brown and Omar Fawzi. Scrambling speed of random quantum circuits. *arXiv preprint arXiv:1210.6644*, 2012.

[19] Stephen H Shenker and Douglas Stanford. Black holes and the butterfly effect. *Journal of High Energy Physics*, 2014(3):1–25, 2014.

[20] Aram Harrow and Saeed Mehraban. Approximate unitary $t$-designs by short random quantum circuits using nearest-neighbor and long-range gates. *arXiv preprint arXiv:1809.06957*, 2018.

[21] Nicholas Hunter-Jones. Unitary designs from statistical mechanics in random quantum circuits. *arXiv preprint arXiv:1905.12053*, 2019.

[22] Jianxin Chen, Dawei Ding, and Cupjin Huang. Randomized benchmarking beyond groups. *arXiv:2203.12703 [quant-ph]*, Mar 2022. arXiv: 2203.12703.

[23] Philip Krantz, Morten Kjaergaard, Fei Yan, Terry P Orlando, Simon Gustavsson, and William D Oliver. A quantum engineer’s guide to superconducting qubits. *Applied Physics Reviews*, 6(2):021318, 2019.

[24] Colin D Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M Sage. Trapped-ion quantum computing: Progress and challenges. *Applied Physics Reviews*, 6(2):021314, 2019.
[25] Alexander M Dalzell, Nicholas Hunter-Jones, and Fernando GSL Brandão. Random quantum circuits transform local noise into global white noise. *arXiv preprint arXiv:2111.14907*, 2021.

[26] Jay M Gambetta, Antonio D Córcoles, Seth T Merkel, Blake R Johnson, John A Smolin, Jerry M Chow, Colm A Ryan, Chad Rigetti, Stefano Poletto, Thomas A Ohki, et al. Characterization of addressability by simultaneous randomized benchmarking. *Physical review letters*, 109(24):240504, 2012.

[27] Andrew W Cross, Lev S Bishop, Sarah Sheldon, Paul D Nation, and Jay M Gambetta. Validating quantum computers using randomized model circuits. *Physical Review A*, 100(3):032328, 2019.

[28] Alexander Erhard, Joel J Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban A Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt. Characterizing large-scale quantum computers via cycle benchmarking. *Nature communications*, 10(1):1–7, 2019.

[29] Robin Harper, Wenjun Yu, and Steven T Flammia. Fast estimation of sparse quantum noise. *PRX Quantum*, 2(1):010322, 2021.

[30] Steven T Flammia. Averaged circuit eigenvalue sampling. *arXiv preprint arXiv:2108.05803*, 2021.

[31] Yihong Zhang, Wenjun Yu, Pei Zeng, Guoding Liu, and Xiongfeng Ma. Scalable fast benchmarking for individual quantum gates with local twirling. *arXiv preprint arXiv:2203.10320*, 2022.

[32] Jonas Helsen, Sepehr Nezami, Matthew Reagor, and Michael Walter. Matchgate benchmarking: Scalable benchmarking of a continuous family of many-qubit gates. *Quantum*, 6:657, 2022.

[33] Timothy J Proctor, Arnaud Carignan-Dugas, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young. Direct randomized benchmarking for multiqubit devices. *Physical review letters*, 123(3):030503, 2019.

[34] Timothy Proctor, Stefan Seritan, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young. Scalable randomized benchmarking of quantum computers using mirror circuits. *arXiv preprint arXiv:2112.09853*, 2021.

[35] Richard Cleve, Debbie Leung, Li Liu, and Chunhao Wang. Near-linear constructions of exact unitary 2-designs. *Quantum Information & Computation*, 16(9-10):721–756, 2016.

[36] R Oliveira, OCO Dahlsten, and MB Plenio. Generic entanglement can be generated efficiently. *Physical review letters*, 98(13):130502, 2007.

[37] Marko Žnidarič. Exact convergence times for generation of random bipartite entanglement. *Physical Review A*, 78(3):032324, 2008.

[38] Winton Brown and Omar Fawzi. Decoupling with random quantum circuits. *Communications in mathematical physics*, 340(3):867–900, 2015.
[39] Fernando GSL Brandao, Aram W Harrow, and Michal Horodecki. Local random quantum circuits are approximate polynomial-designs. *Communications in Mathematical Physics*, 346(2):397–434, 2016.

[40] Matthew B Hastings and Jeongwan Haah. Dynamically generated logical qubits. *Quantum*, 5:564, 2021.

[41] Richard Jozsa and Akimasa Miyake. Matchgates and classical simulation of quantum circuits. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2100):3089–3106, 2008.

[42] Héctor J. García, Igor L. Markov, and Andrew W. Cross. On the geometry of stabilizer states. *Quantum Info. Comput.*, 14(7 & 8):683–720, may 2014.

[43] Laurent Saloff-Coste. Random walks on finite groups. In *Probability on discrete structures*, pages 263–346. Springer, 2004.