A Theory of Direct Randomized Benchmarking

Anthony M. Polloreno[1,2], Arnaud Carignan-Dugas[3,4], Jordan Hines[5,2], Robin Blume-Kohout[2], Kevin Young[2], and Timothy Proctor[2]

[1]JILA, NIST and Department of Physics, University of Colorado, 440 UCB, Boulder, Colorado 80309, USA
[2]Quantum Performance Laboratory, Sandia National Laboratories, Livermore, CA 94550 and Albuquerque, NM 87185
[3]Institute for Quantum Computing and the Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[4]Keysight Technologies Canada, Kanata, ON K2K 2W5, Canada
[5]Department of Physics, University of California, Berkeley, CA 94720

Randomized benchmarking (RB) protocols are widely used to measure an average error rate for a set of quantum logic gates. However, the standard version of RB is limited because it only benchmarks a processor's native gates indirectly, by using them in composite n-qubit Clifford gates. Standard RB's reliance on n-qubit Clifford gates restricts it to the few-qubit regime, because the fidelity of a typical composite n-qubit Clifford gate decreases rapidly with increasing n. Furthermore, although standard RB is often used to infer the error rate of native gates, by rescaling standard RB's error per Clifford to an error per native gate, this is an unreliable extrapolation. Direct RB is a method that addresses these limitations of standard RB, by directly benchmarking a customizable gate set, such as a processor's native gates. Here we provide a detailed introduction to direct RB, we discuss how to design direct RB experiments, and we present two complementary theories for direct RB. The first of these theories uses the concept of error propagation or scrambling in random circuits to show that direct RB is reliable for gates that experience stochastic Pauli errors. We prove that the direct RB decay is a single exponential, and that the decay rate is equal to the average infidelity of the benchmarked gates, under broad circumstances. This theory shows that group twirling is not required for reliable RB. Our second theory proves that direct RB is reliable for gates that experience general gate-dependent Markovian errors, using similar techniques to contemporary theories for standard RB. Our two theories for direct RB have complementary regimes of applicability, and they provide complementary perspectives on why direct RB works. Together these theories provide comprehensive guarantees on the reliability of direct RB.

1 Introduction

Reliable, efficient, and flexible methods for benchmarking quantum computers are becoming increasingly important as 5-50+ qubit processors become commonplace. Isolated qubits or coupled pairs can be studied in detail with tomographic methods [1, 2, 3], but tomography of general n-qubit processes requires resources that scale exponentially with n. Randomized benchmarking

Anthony M. Polloreno: ampolloreno@gmail.com
Timothy Proctor: tjproct@sandia.gov

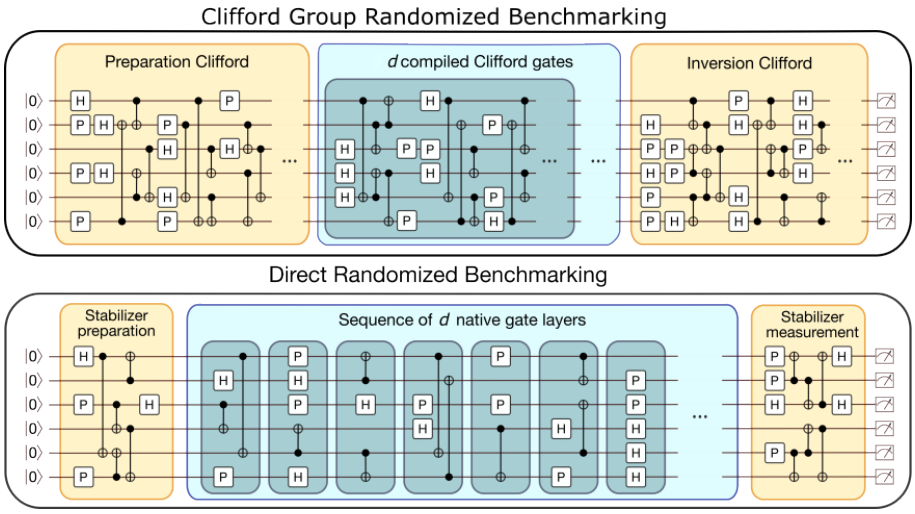

Figure 1: An illustration (adapted from [20]) comparing the varied-length random circuits used in the standard RB protocol of Magesan *et al.* [7, 8] ("Clifford group RB") and the streamlined direct RB protocol that was introduced in [20] and that we investigate further here. Direct RB can be implemented on almost any n-qubit gate set (a set of circuit "layers" or "cycles") that generates a unitary 2-design. The case shown here is when that gate set is a set of Clifford gates (CNOT, Hadamard, and the phase gate) that generate the n-qubit Clifford group. The circuits of direct RB can be shallower than those of Clifford group RB (e.g., consider the case of $d = 0$ in each circuit), allowing for RB on more qubits.

(RB) methods [4, 5, 6, 7, 8, 9, 10, 11, 12, 7, 8, 13, 9, 10, 11, 12, 14, 15, 16, 17, 18] were introduced partly to avoid the scaling problems that afflict tomography. RB avoids these scaling problems because, in the most commonly used RB protocols [7, 8, 13], both the necessary number of experiments [19] and the complexity of the data analysis [8] are independent of n. RB methods efficiently estimate a single figure of merit—the average infidelity of a gate set—by (i) running random circuits of varied depths (d) that should implement an identity operation, (ii) observing the probability that the input is successfully returned (S_d) versus d, and then (iii) fitting this data to an exponential. The decay rate of this exponential (r) is an estimate of the gates' average infidelity.

However, many RB methods have a different scaling problem, introduced by a *gate compilation* step. Most RB protocols benchmark an n-qubit gate set that forms a group [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], so that they can use group twirls as their theoretical foundations. But a quantum processor's native operations do not normally form a group. Instead, subsets of a processor's native gates can be used to *generate* a variety of different groups (e.g., the unitary group, the Clifford group, the Pauli group, etc). This is a problem for implementing group-based RB methods because, for many n-qubit groups, the number of basic gates required to implement a typical group element is very large, meaning that many group elements can only be implemented with low fidelity.

The standard RB protocol—which we will call "Clifford group RB"—benchmarks the n-qubit Clifford group [7, 8], estimating an error rate (r) that corresponds closely to the average infidelity of a gate from the n-qubit Clifford group [21, 22]. This protocol runs random sequences of $d + 2$ composite gates from that group (see the upper circuit in Fig. 1), with a variable benchmark depth (d, with $d \geq 0$). The first $d + 1$ gates are sampled uniformly from the Clifford group, and the last gate is selected to invert the proceeding $d + 1$ gates. For a single qubit there are only 24 Clifford gates, so each Clifford gate can typically be compiled into a short sequence of native gates. But the Clifford group grows quickly with the number of qubits n: there are $2^{O(n^2)}$ n-qubit Clifford gates [23, 24], and a single typical n-qubit Clifford gate can be compiled into no fewer than $O(n^2/\log n)$

one- and two-qubit gates [25, 26, 27]. This means that the fidelity of a random Clifford gate degrades rapidly with n for a fixed quality of one- and two-qubit gates, i.e., $r \to 1$ quickly with increasing n. This makes it infeasible to reliably estimate Clifford group RB's error rate r on many qubits—even with very high fidelity one- and two-qubit gates—because as n increases even the shortest Clifford group RB circuits ($d = 0$, corresponding to a random n-qubit Clifford gate and its inverse) typically have such small success probabilities that they cannot be distinguished from $1/2^n$ with a reasonable number of circuit repetitions (see Fig. 2). Indeed, one- and two-qubit Clifford group RB has been widely implemented [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41] but we are aware of only two publications presenting Clifford group RB experiments on more than two qubits [42, 43], and none presenting Clifford group RB experiments on six or more qubits.

An alternative to Clifford group RB is to benchmark a gate set that is a smaller group [9, 12, 10, 11, 14, 16], enabling group elements to be implemented with fewer native gates. The gate compilation overhead can then be reduced from Clifford group RB, but often at a cost—the data analysis and the experiments become more complex. The simple behavior of Clifford group RB, i.e., the exponential decay of its success probability data, is guaranteed by a particular property of the Clifford group: conjugation of an error channel by a uniformly random Clifford group element "twirls" that channel into an n-qubit depolarizing channel. This follows because the Clifford group is a unitary 2-design [44, 45, 46] (or, equivalently, because the superoperator representation of Clifford gates consists of the direct sum of two irreducible representations of the Clifford group). The less like a unitary 2-design a group is (i.e., the more irreducible representations of the group its superoperator representation decomposes into) the weaker the effect of twirling over it is, and so the more complex the RB decay curve can be [9, 12, 10, 11, 14, 16] (see [16] for a theory of RB over general groups). For general groups, the RB decay curve is a sum of many exponentials, rather than the single exponential of Clifford group RB.

Another consequence of the gate compilation required in group-based RB protocols is that they measure an error *per compiled group element*. There are infinitely many ways to compile each group element into native gates, and the observed error rate depends strongly on which group compilations is used. This is desirable if we aim to estimate the fidelity with which gates in that group can be implemented. But normally group-based RB error rates are used as a proxy for native gate performance [37, 38, 39, 42]. In particular, it is common to estimate a native gate error rate from a Clifford group RB error rate, e.g., by dividing Clifford group RB's r by the average circuit size of a compiled gate [37, 38, 39] or by the average number of two-qubit gates in a compiled gate [42] (possibly after correcting for the estimated contribution of one-qubit gates). However, these rescaling methods have little theoretical justification [47, 20]: in general, the resultant error rate is not a reliable estimate of the average infidelity of the native gates.

One solution to the limitations of Clifford group RB is *direct randomized benchmarking* [20]. Direct RB is a technique that is designed to retain the conceptual and experimental simplicity and efficiency of Clifford group RB while enabling direct benchmarking of a processor's native gates. Like all RB protocols, direct RB utilizes random circuits of variable length. Furthermore, like Clifford group RB, direct RB's data analysis simply consists of (1) fitting success probability data to a single exponential decay, and (2) rescaling the fit decay rate to obtain an estimate of the benchmarked gates' average infidelity. But, unlike Clifford group RB, the varied-length portion of direct RB circuits can consist of randomly sampled layers of a processor's native gates, rather than compiled group operations (see the lower circuit in Fig. 1). Direct RB therefore enables estimating the average infidelity of native gate layers. The only restriction on the gate set that can be benchmarked with direct RB is that it must *generate* a group that is a unitary 2-design (e.g., the n-qubit Clifford group).

Direct RB addresses the main limitations of Clifford group RB, but it has been lacking a theory that proves that it is reliable. Those existing theories for RB that are applicable to direct RB

[16, 20, 48] provide few guarantees on direct RB's behavior. In this paper we provide a detailed introduction to direct RB and two complementary theories of direct RB. These theories show that direct RB is reliable—the direct RB success probability follows an exponential decay and the direct RB error rate is the infidelity of the benchmarked gates—and explain why direct RB works. Shortly before the completion of this work, Heinrich *et al.* [49] presented a theory for *filtered RB*, which is a technique that enables RB of general gate sets using random circuits containing i.d.d. gates from that set, by using a classical post-processing of the data which they call *filtering*. That work presents ideas and methods that are distinct from our work here (e.g., they analyze a different kind of RB) using related techniques.

This paper's main results and structure are as follows. In Section 2 we introduce our notation and review the necessary background material. In Section 3 we provide a detailed definition and introduction to the direct RB protocol, and we compare it to other RB protocols for benchmarking native gate sets [6, 50, 48]. The definition for direct RB presented here generalizes that in [20]. This section includes examples that illustrate how direct RB can be used to benchmark a variety of physically relevant gate sets—including universal gate sets on few qubits. We also present a simulation of direct RB that shows (approximately) how many qubits direct RB can *feasibly* be used to benchmark, given the magnitude of current gate error rates. This limit depends on a variety of factors, but for realistic gate and readout error rates (between 0.1% and 1%) our numerical results suggest that it is feasible to benchmark around 10-15 qubits with direct RB (although recent extensions of direct RB [43, 51, 52], not discussed in detail here, remain feasible well beyond this limit).

In Section 4 we present a theory for direct RB on gates that experience stochastic Pauli errors, which is the relevant error model for gates that have undergone Pauli frame randomization [53, 54] or randomized compilation [55, 56]. This theory provides a physically intuitive underpinning for direct RB, by using the intuitive concepts of error propagation and scrambling in random circuits. Interestingly, this theory shows that group twirling is not necessary for reliable RB—direct RB can be reliable even when the benchmarked gate set is very far from approximating a unitary 2-design, and these insights are broadly applicable to benchmarks that use random circuits. This enables us to show that direct RB is reliable even in the $n \gg 1$ regime, where short sequences of layers of native gates cannot approximate a unitary 2-design.

One of the contributions of this paper is a theory for direct RB with general gate-dependent Markovian errors that is similar in nature to the group twirling (and Fourier analysis) theories for group-based RB methods. This theory uses the concept of a *sequence-asymptotic unitary 2-design*, which we introduce in Section 5. A sequence-asymptotic unitary 2-design is any gate set for which length d random sequences of gates from that set create a distribution over unitaries that converges to a unitary 2-design as $d \to \infty$. Sequence-asymptotic unitary 2-design are closely related to various existing notions of approximate or asymptotic unitary designs and the general theory of scrambling in random circuits [45, 44, 57, 58, 59, 60, 61, 62, 63, 49]. However, to our knowledge, the particular concept and technical results that we require for our theory of direct RB do not appear in the literature.

In Section 6 we present our theory for direct RB under general gate-dependent Markovian errors, which is based on twirling superchannels (or Fourier transforms on groups). First, we show that the depth-d direct RB success probability can be computed exactly by taking the d^{th} power of a superchannel ($\mathscr{R}$) constructed from the permutation matrix representation of the group generated by the benchmarked gate set. Then we show how the direct RB decay can be (approximately) rewritten in terms of a smaller superchannel ($\mathscr{L}$) that is constructed from the superoperator representation of the generated group. These results correspond to equivalent theories for Clifford group RB presented in [21, 22, 64, 65]. We then use the spectral decomposition of these superchannels, together with the theory of sequence-asymptotic unitary 2-designs, to show that the direct RB de-

	Theory in Section 4	Theory in Section 6
Techniques	Error scrambling Stabilizer state theory	Group Fourier transform Unitary 2-designs
Assumptions	Clifford gates Stochastic Pauli channels $n \gg 1$ $\epsilon \ll 1$ ϵ constant as n increases	Any gates General Markovian errors Any n $\epsilon \ll 1$ ϵ decreasing as n increases
Result	$r_\Omega \approx \epsilon_\Omega$	$r_\Omega \approx \epsilon_\Omega$ with caveats

Table 1: A summary of the two complementary theories for direct RB that we present. The "techniques" row concisely describes the mathematical techniques used within each theory. The "assumptions" row describes the key assumptions made in each theory. Here n denotes the number of qubits, and ϵ the error per n-qubit gate (a.k.a. circuit layer) in the gate set being benchmarked. Both theories require that ϵ is small ($\epsilon \ll 1$), but our theory that is applicable to general Markovian errors has a much more stringent condition on how small ϵ is as a function of n. The "result" row summarizes the key result of each theory. Both theories show that direct RB's error rate (r_Ω) is approximately equal to the weighted average of the benchmarked gates' infidelities (ϵ_Ω). However, in the case of general Markovian errors (the theory of Section 6), this result has subtle caveats that are related to the difficulty in defining a gate's fidelity in the presence of general Markovian errors.

cay is approximately a singe exponential. Finally, we use the $\mathscr{L}$ superchannel to show that the direct RB error rate is equal to a gauge-invariant version of the mean fidelity of the benchmarked gates, building on Wallman's derivation of an equivalent result for Clifford group RB [22].

Our two theories of direct RB are complementary, providing different perspectives as well as different regimes of applicability. We summarize these differences in Table 1. Our superchannel-based theory (Section 6) is more mathematically rigorous than our scrambling-based theory (Section 4), and it applies to a more general class of errors. However, our superchannel-based theory relies on an approximation whose size increases with n, whereas our scrambling-based theory shows that direct RB is reliable even when $n \gg 1$. Therefore, together these theories provide comprehensive evidence for the reliability of direct RB.

2 Definitions

In this section we introduce our notation and review the necessary background material.

2.1 Gates, gate sets, and circuits

An n-qubit gate G is a physical operation associated with an instruction to ideally implement a particular unitary evolution on n qubits (we use the term "gate" for consistency with the RB literature—an n-qubit gate is also often referred to as a "layer" or a "cycle"). We denote the unitary corresponding to G by $U(G) \in \mathrm{SU}(2^n)$. It will also often be convenient to use the superoperator representation of a unitary, so we define $\mathcal{G}(G)$ to be the linear map

$$\mathcal{G}(G)[\rho] := U(G)\rho U(G)^\dagger, \tag{1}$$

where ρ is an n-qubit density operator. We consider a gate G to be entirely defined by the unitary $U(G)$. RB methods are agnostic about how an n-qubit gate is implemented, except that an attempt must be made to faithfully implement the unitary it defines. We will use $\mathbb{G}$ to denote a set of

n-qubit gates, which need not be a finite set, and, slightly abusing notation, we will also use $\mathbb{G}$ to denote the corresponding sets of unitaries (i.e., $\{U(G) \mid G \in \mathbb{G}\}$ and superoperators (i.e., $\{\mathcal{G}(G) \mid G \in \mathbb{G}\}$) with the meaning implied by the context.

A quantum circuit C over a n-qubit gate set $\mathbb{G}$ is a sequence of $d \geq 0$ gates from $\mathbb{G}$. We will write this as

$$C = G_d G_{d-1} \cdots G_2 G_1, \tag{2}$$

where each $G_i \in \mathbb{G}$, and we use a convention where the circuit is read from right to left. The circuit C is an instruction to apply its constituent gates, G_1, G_2, ..., in sequence, and it implements the unitary

$$U(C) = U(G_d)U(G_{d-1}) \cdots U(G_2)U(G_1). \tag{3}$$

In direct RB, and most other RB methods, multiple gates in a circuit must *not* be combined or compiled together by implementing a physical operation that enacts their composite unitary (there are "barriers" between gates).

2.2 Fidelity

In our theory for direct RB we will show how the direct RB error rate is related to the fidelity of the benchmarked gates. There are two commonly used version of a gate's fidelity: average gate fidelity and entanglement fidelity, defined below. Throughout this work, we will use the Markovian model for errors, whereby the imperfect implementation of a gate G is represented by a complete positive and trace preserving (CPTP) superoperator $\tilde{\mathcal{G}}(G)$ [so, for low-error gates $\tilde{\mathcal{G}}(G) \approx \mathcal{G}(G)$]. The *average gate fidelity* of $\tilde{\mathcal{G}}(G)$ to $\mathcal{G}(G)$ is defined by

$$F_{\text{A}}(\tilde{\mathcal{G}}, \mathcal{G}) := \int d\psi \ \text{Tr}\left\{\tilde{\mathcal{G}}[|\psi\rangle\langle\psi|]\, \mathcal{G}[|\psi\rangle\langle\psi|]\right\}, \tag{4}$$

where $d\psi$ is the normalized Haar measure [66], and where here we have dropped the dependence of $\mathcal{G}$ and $\tilde{\mathcal{G}}$ on G for brevity (we also do this elsewhere when convenient). The *entanglement fidelity* is defined by

$$F_{\text{E}}(\tilde{\mathcal{G}}, \mathcal{G}) := \langle\varphi|(\mathbb{1} \otimes \Lambda)[|\varphi\rangle\langle\varphi|]|\varphi\rangle, \tag{5}$$

where φ is any maximally entangled state [66] and

$$\Lambda(G) = \tilde{\mathcal{G}}(G)\mathcal{G}(G)^{-1}. \tag{6}$$

The corresponding infidelities, *average gate infidelity* (ϵ_{A}) and *entanglement infidelity* (ϵ_{E}), are defined by

$$\epsilon_{\text{A/E}}(\tilde{\mathcal{G}}, \mathcal{G}) = 1 - F_{\text{A/E}}(\tilde{\mathcal{G}}, \mathcal{G}). \tag{7}$$

These infidelities are related by [66]

$$\epsilon_{\text{E}} = \frac{2^n + 1}{2^n}\epsilon_{\text{A}}. \tag{8}$$

The average gate [in]fidelity is more commonly used in the RB literature, but in this work we use the entanglement [in]fidelity. Typically we drop the subscript, letting $\epsilon \equiv \epsilon_{\text{E}}$. Note that a gate's [in]fidelity is not a "gauge-invariant" property of a gate set [21], meaning that it is not technically measurable: we discuss this subtle point when presenting our theory for direct RB with general gate-dependent Markovian errors (Section 6).

Randomized benchmarking methods are typically designed to measure the mean infidelity of a set of gates. Direct RB is designed to measure

$$\epsilon_\Omega = \sum_{G\in\mathbb{G}} \Omega(G)\epsilon\left[\tilde{\mathcal{G}}(G), \mathcal{G}(G)\right], \tag{9}$$

where Ω is a probability distribution over the gate set $\mathbb{G}$. Note that, except where otherwise stated, $\mathbb{G}$ does not need to be a finite set, i.e., it can include gates with continuous parameters. The $\sum_{G\in\mathbb{G}}\Omega(G)$ notation is a shorthand for integrating and/or summing over $\mathbb{G}$ according to the measure $\Omega(G)$. When gates experience only stochastic Pauli errors, then ϵ_Ω is equal to the probability that a Pauli error occurs on a gate sampled from Ω.

3 Direct randomized benchmarking

In this section we first define the direct RB protocol and explain what its purpose is (Section 3.1). Direct RB has flexible, user-specified components, so we then provide guidance on how to choose these aspects of direct RB experiments (Sections 3.3-3.5). We then explain why the direct RB protocol is defined the way it is, and we discuss how it differs from Clifford group RB (Sections 3.6-3.8). Finally, we compare direct RB to other RB protocols for directly benchmarking native gate sets (Section 3.10).

3.1 The direct RB protocol

We now define the n-qubit direct RB protocol, and explain what it aims to measure. The definition given here is more general than that in [20]. For example, in [20], direct RB is defined for gate sets containing only Clifford gates. Our definition here permits non-Clifford gates. Note that our definition for direct RB applies to gates on a set of $n \geq 1$ qubits. Generalizing to $n \geq 1$ qudits of arbitrary dimension is conceptually simple, as is the case with Clifford group RB, but we do not pursue this here. In addition to parameters that set the number of samples (which exist in all RB protocols), direct RB on n qubits has two flexible, user-specified inputs:

1. A set of n-qubit gates (a.k.a., layers or cycles) to benchmark ($\mathbb{G}$).

2. A sampling distribution Ω over the gate set $\mathbb{G}$.

We denote direct RB of the gate set $\mathbb{G}$ with the sampling distribution Ω by DRB$(\mathbb{G}, \Omega)$. We call $\mathbb{G}$ a "gate set" for consistency with common RB terminology, but note that $\mathbb{G}$ contains n-qubit gates—i.e., "circuit layers" or "cycles"—not one- and two-qubit gates. All n-qubit gate sets generate some n-qubit group, or a dense subset of some group. This group plays an important role in the direct RB protocol, and we denote it by $\mathbb{C}$ (for brevity, for a gate set $\mathbb{G}$ that only generates a dense subset of a group $\mathbb{C}$ we also refer to $\mathbb{G}$ as generating $\mathbb{C}$). In most of our examples this group will be the n-qubit Clifford group, but this is not required (the conditions required of $\mathbb{C}$ are stated in Section 3.4, which include that it is a unitary 2-design).

Having introduced $\mathbb{G}$ and Ω, we are now ready to define the direct RB protocol. DRB$(\mathbb{G}, \Omega)$ is a protocol for estimating ϵ_Ω [see Eq. (9)] and it consists of the following procedure:

1. **Sample the circuits.** For each d in some set of user-specified *benchmark depths* all satisfying $d \geq 0$, and for $k = 1, 2, \ldots, K_d$ for some user-specified $K_d \geq 1$:

    1.1. Choose a target output n-bit string $s_{d,k}$ for this circuit. This can be set to the all-zeros bit string (as in the conventional implementation of Clifford group RB [7, 8]), or it can be chosen uniformly at random (our recommendation).

1.2. Sample a group element $F_{\rm sp}$ uniformly from $\mathbb{C}$ (the group generated by $\mathbb{G}$), and then find a circuit $C_{\rm sp}$ that creates the state

$$|\psi\rangle = U(F_{\rm sp})|0\rangle^{\otimes n}, \tag{10}$$

from $|0\rangle^{\otimes n}$ [$U(\cdot)$ maps a gate or circuit to the unitary it implements—see Section 2]. That is, find a circuit $C_{\rm sp}$ that satisfies

$$U(C_{\rm sp})|0\rangle^{\otimes n} = U(F_{\rm sp})|0\rangle^{\otimes n}, \tag{11}$$

meaning that $C_{\rm sp}$ is only required to faithfully implement $U(F_{\rm sp})$'s action on $|0\rangle^{\otimes n}$. The circuit $C_{\rm sp}$ can contain any gates, including gates that are not in $\mathbb{G}$, and it is ideally chosen to maximize the fidelity with which ψ is produced. The circuit $C_{\rm sp}$ is the first part of the sampled direct RB circuit, and we refer to it as the circuit's state preparation subcircuit.

1.3. Independently sample d gates, $G_1, G_2, \ldots, G_d$, from Ω (the user-specified distribution over $\mathbb{G}$). The circuit $G_d \cdots G_2 G_1$ is the next part of the sampled direct RB circuit, and we refer to it as the circuit's "core".

1.4. Find a circuit $C_{\rm mp}$ that, when applied after the two parts of the circuit sampled so far, maps the qubits to $|s_{d,k}\rangle$, and append it to the circuit sampled so far. That is, $C_{\rm mp}$ is a circuit that satisfies

$$U(C_{\rm mp}\, G_d \cdots G_2\, G_1\, C_{\rm sp})|0\rangle^{\otimes n} = |s_{d,k}\rangle. \tag{12}$$

The circuit $C_{\rm mp}$ is the final part of the sampled direct RB circuit. As with the circuit $C_{\rm sp}$, $C_{\rm mp}$ can contain any gates, and we refer to it as the circuit's measurement preparation subcircuit.

The sampling procedure of 1.1-1.4 creates the circuit

$$C_{d,k} = C_{\rm mp} G_d \cdots G_2 G_1 C_{\rm sp}, \tag{13}$$

that, if run on a perfect quantum computer, always outputs the bit string $s_{d,k}$. That is, $|\langle s_{d,k}|U(C_{d,k})|0\rangle^{\otimes n}|^2 = 1$.

2. **Run the circuits.** Run each of the sampled circuits N times, for some user-specified $N \geq 1$. Estimate each circuit's success probability ($S_{d,k}$), as the frequency that the circuit $C_{d,k}$ returns its success bit string $s_{d,k}$ (we denote this estimate by $\hat{S}_{d,k}$). Note that the direct RB protocol does not specify the order that the circuits are run, but this is important if there could be significant drift in the system over the time period of the entire experiment [67, 68, 69]. We recommend "rastering" [70] if possible—meaning looping through all of the circuits N times, running each circuit once in each loop—as this facilitates detecting the presence of drift and a time-resolved RB analysis [70] (these analyses are not discussed further herein). If rastering is not possible then the order that circuits are run should be randomized, as this will typically reduce the impact of drift on the results.

3. **Analyze the data.** Fit the estimated average probability of success, $\hat{S}_d = \sum_k \hat{S}_{d,k}/K_d$, versus benchmark depth (d) to the exponential decay function

$$S_d = A + Bp^d, \tag{14}$$

where A, B and p are fit parameters (this is the same as the fit function used in Clifford group RB). If the success bit strings were sampled uniformly then fix $A = 1/2^n$. The estimate of the DRB$(\mathbb{G}, \Omega)$ error rate of the gates ($\hat{r}_\Omega$) is then defined to be

$$\hat{r}_\Omega = \frac{(4^n - 1)}{4^n}(1 - \hat{p}). \tag{15}$$

where $\hat{p}$ is the fit value of p.

Our definition for the direct RB procedure specifies fitting the data to the standard exponential decay function of Eq. (14), and so DRB$(\mathbb{G}, \Omega)$ is only well-behaved if the direct RB data has approximately this form. However, it is not clear *a priori* that the direct RB average success probability will decay exponentially. Moreover, even if the decay is an exponential it is not obvious what the direct RB error rate (r_Ω) measures. Proving that $\hat{S}_d \approx A + Bp^d$ under broad conditions, providing an intuitive explanation for why $\hat{S}_d$ decays exponentially, and then explaining what r_Ω quantifies are three of the main aims of this paper. In particular, we will show that $r_\Omega \approx \epsilon_\Omega$.

3.2 A simple numerical demonstration of direct RB

We now demonstrate that direct RB works correctly—the decay is an exponential and $r_\Omega \approx \epsilon_\Omega$—in two simple scenarios. First, consider the case of gate-independent global depolarizing noise, and assume perfect state preparation and measurement sub-circuits. In this model, after each n-qubit gate the n qubits are mapped to the completely mixed state with some probability $1 - \lambda$. This is arguably the simplest possible error model, and we would expect the error rate measured by any well-formed RB protocol to be the infidelity of this global depolarizing channel, which is $\epsilon = (4^n - 1)(1 - \lambda)/4^n$. An explicit calculation confirms that this is the case here:

$$S_d = 1/2^n + (1 - 1/2^n)\lambda^d, \tag{16}$$

which implies that $p = \lambda$ and $r_\Omega = (4^n - 1)(1 - \lambda)/4^n$.

Global depolarizing errors are physically unrealistic, so we also consider another error model that is more realistic but that is still simple to understand: gate-independent *local* depolarization. We simulated n-qubit direct RB for a hypothetical n-qubit processor (with $n = 2, 4, 6, \ldots, 14$) whereby after each n-qubit gate is applied every qubit experiences independent uniform depolarization with an error rate of 0.1%, i.e., the infidelity of each single-qubit depolarizing channel is $\epsilon = 0.001$. This error model has the convenient property that we can easily compute ϵ_Ω. The (entanglement) fidelity of a tensor product of n error channels $\mathcal{E}_1, \ldots, \mathcal{E}_n$ is the product of $\mathcal{E}_1$ to $\mathcal{E}_n$'s fidelities, i.e.,

$$F(\mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n, I \otimes \cdots \otimes I) = F(\mathcal{E}_1, I) \cdots F(\mathcal{E}_n, I), \tag{17}$$

so ϵ_Ω is given by $\epsilon_\Omega = 1 - (1 - 0.001)^n \approx n \times 0.001$. Note that, because the error rates are gate-independent, ϵ_Ω is again independent of the sampling distribution Ω in this example. Figure 2 (upper plot) shows S_d versus d (black circles) as well as the fit exponential (solid lines) and the estimates for r_Ω (in the legend), for each n. We observe that S_d decays exponentially, and that $\hat{r}_\Omega \approx \epsilon_\Omega$. We specify the gate set and sampling distribution used in this simulation in Example 2 of Section 3.4.

These numerical results also enable a heuristic assessment of how many qubits can feasibly be benchmarked using direct RB, given the approximate size of current gate error rates (and our suboptimal compilers for creating stabilizer states and measurements). We observe that, in these simulations, S_d's value at $d = 0$ is significantly non-zero for even 14 qubits. Readout errors (which are not included in this simulation) would suppress this intercept further, but readout error

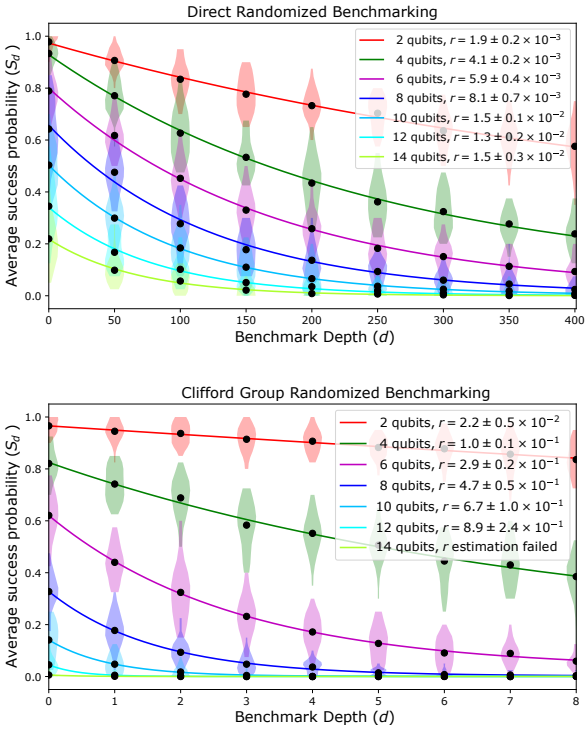

Figure 2: Simulated n-qubit direct RB and Clifford group RB for $n = 2, 4, 6, \ldots, 14$. Here, direct RB is benchmarking n-qubit gates that consist of a single layer of parallel one- and two-qubit gates. Clifford group RB is benchmarking the n-qubit Clifford group generated out of these layers. In this simulation, an imperfect n-qubit layer is modeled by the perfect unitary followed by independent uniform depolarization on each qubit at a rate of 0.1%. The points and violin plots are the means and the distributions of the estimated circuit success probabilities versus benchmark depth (d), respectively, and the lines are fits of the means to $S_d = A + Bp^d$. The estimated RB error rates ($\hat{r}$), which are reported in the legends, are obtained from the fit decay rates, using Eq. (15). We observe that the direct RB average circuit success probabilities (S_d) decay exponentially for all n, and that the direct RB error rate is given by $\hat{r} \approx 1 - (1 - 0.001)^n \approx n \times 0.001$. Direct RB is therefore accurately estimating the error rate of the n-qubit gates it is benchmarking. The Clifford group RB error rate grows very quickly with n—as it is benchmarking the n-qubit Clifford group, whose elements must be compiled into the available layers of parallel one- and two-qubit gates. The Clifford group RB $d = 0$ intercept quickly decays to approximately $1/2^n$ as n increases, at which point r is close to 100% and it cannot be estimated to reasonable precision with a practical amount of data. This demonstrates that direct RB is feasible on substantially more qubits than standard Clifford group RB.

of a similar size to the gate errors here (0.1%) would have negligible effect on S_0. Higher error rates would limit direct RB to fewer qubits. We can estimate the largest feasible number of qubits using the numerical results in Section 3.7, where we compute the number of two-qubit gates in the circuits created by our (suboptimal) algorithm for generating random n-qubit stabilizer states. These results suggest that with, e.g., a two-qubit gate error rate of 0.5% it would be feasible to benchmark around 12 qubits using direct RB. In summary, our assessment is that around 10-15 qubit could be benchmarked using direct RB for gate and readout errors in the range of 0.1%-1%. Furthermore, values for a given set of error rates can be easily computed using simulations. We note, however, that more scalable versions of direct RB [43, 51, 52] are, in our experience, likely to be more useful in practice beyond around 8-10 qubits.

3.3 Direct RB's standard sampling parameters

The direct RB protocol has three user-specified parameters in common with all RB protocols, which we now briefly discuss. These are the user-specified benchmark depths (the values for d), the number of repetitions of each circuit (N), and the number of randomly sampled circuits at each benchmark depth (K_d). These parameters predominantly control purely statistical aspects of the experiment—specifically, the precision with which the direct RB experiment estimates the underlying $N, K_d \to \infty$ direct RB error rate. How these parameters control the statistical uncertainty in an estimated RB error rate, and how to choose these parameters, has been studied in some detail for Clifford group RB [8, 46, 47, 71, 19, 72]. Much of this work can be applied to direct RB. We will therefore not discuss these parameters in detail herein. Instead, we will provide some simple numerical evidence that the direct RB error rate can be estimated from reasonable amounts of data. The simulated direct RB experiments of Fig. 2 use a practical amount of data: we used 9 benchmarking depths, $K_d = 30$, and $N = 40$, for a total of approximately 10^4 samples for the direct RB instance at each value of n. Furthermore, the estimated direct RB error rates have low uncertainty (the uncertainties reported in the legend are 2σ and they are estimated using a standard bootstrap). Throughout the remainder of this paper we are predominantly interested in studying the behaviour of direct RB in the limit of $N, K_d \to \infty$; we denote the average circuit success probability and the direct RB error rate in this limit by S_d and r_Ω, respectively.

3.4 Selecting the gate set to benchmark

One of the defining characteristics of direct RB is that the user specifies the n-qubit gate set ($\mathbb{G}$) that is to be benchmarked. This is in contrast with Clifford group RB, where the benchmarked gate set is necessarily the n-qubit Clifford group—or another group that is a unitary 2-design. However there is not total freedom in choosing $\mathbb{G}$ in direct RB. The direct RB gate set has to satisfy the following conditions, some of which depend on the performance of the processor that is to be benchmarked:

- The group $\mathbb{C} \subseteq \mathrm{SU}(2^n)$ generated by $\mathbb{G}$ must be a unitary 2-design over $\mathrm{SU}(2^n)$. The requirement that $\mathbb{C}$ is a unitary 2-design can likely be relaxed (with appropriate adaptions to the direct RB experiments and/or data analysis) using techniques from RB for general groups [9, 10, 11, 12, 14, 15, 16, 17]. This is an interesting open area of research, as these generalizations have the potential to enable direct RB on any gate set $\mathbb{G}$. Our formal theory for direct RB (Section 6) relies on a slightly stronger condition on $\mathbb{G}$ than simply generating a unitary 2-design—it requires that $\mathbb{G}$ induces a *sequence-asymptotic unitary 2-design*. The concept of a sequence-asymptotic unitary 2-design is introduced in Section 5, where we state necessary and sufficient conditions for a gate set to induce a sequence-asymptotic unitary 2-design. Here we state a simple condition that is sufficient (but not necessary) for a gate set $\mathbb{G}$ to induce a sequence-asymptotic unitary 2-design: the gate set $\mathbb{G}$ generates a unitary 2-design and contains the identity operation.

- The n qubits to be benchmarked must have sufficiently high fidelity operations for there to exist a circuit to prepare any random state from $\{|\psi\rangle = U(C)|0\rangle^{\otimes n} \mid C \in \mathbb{C}\}$ with non-negligible fidelity. To actually run direct RB, the user also needs an explicit algorithm for finding these circuits. This is required so that errors in the state and measurement preparation subcircuits do not mean that $S_d \approx 1/2^n$ at $d = 0$ (how far above $1/2^n$ it is necessary for S_0 to be depends on how much data the user is willing to collect).

- It must be feasible to sample the direct RB circuits using a classical computer, which requires, e.g., multiplying together arbitrary elements in $\mathbb{C}$, and sampling uniformly from $\mathbb{C}$.

This condition is also required to implement standard RB over $\mathbb{C}$, and it is satisfied if $\mathbb{C}$ is the Clifford group.

None of the three conditions on $\mathbb{G}$, above, require any efficient scaling with the number of qubits n. Instead, it suffices that each condition is satisfied at the values for n of interest. For example, perhaps only benchmarking one- or two-qubit gate sets is of interest (note that most RB experiments are one- or two-qubit Clifford group RB). The conditions also do not require that $\mathbb{G}$ is finite. This is illustrated by our first example, below, of a practically interesting type of gate set that can be benchmarked with direct RB.

Example gate set 1 [$\mathbb{G}_1(n)$]: An n-qubit gate set constructed from all possible combinations of parallel applications of CNOT gates on connected qubits and the one-qubit gates $X(\pi/2)$ and $Z(\theta)$ for $\theta \in (-\pi, \pi]$ [where $X(\theta)$ and $Z(\theta)$ denote rotations by θ around σ_x and σ_z, respectively].

This gate set is universal—i.e., it generates SU(2^n)—and so $\mathbb{G}_1(n)$ does not satisfy all our conditions for $n \gg 1$ [generating a uniformly (Haar) random pure state is not efficient in n]. However, this gate set does satisfy our conditions for small n. For example, generating any unitary in SU(4) requires only three CNOT gates [73, 74]. The particular n-qubit universal gate set given above is just an example, and this reasoning holds for (almost) all few-qubit universal gate sets. Direct RB can therefore be used to benchmark almost any universal gate set over a few qubits.[1] This is arguably substantially simpler than the RB method used by Garion *et al.* [75] to benchmark a two-qubit gate set containing a controlled $Z(\pi/4)$ gate (which is not a Clifford gate). Note that, recently, direct RB of one- and two-qubit universal gate sets has been experimentally demonstrated in [51].

Example gate set 2 [$\mathbb{G}_2(n)$]: An n-qubit gate set constructed from CNOT gates and any set of generators for the single-qubit Clifford group (such as the Hadamard and phase gates).

This is the type of gate set used in the simulations of Fig. 2 (and shown in the schematic of Fig. 1). In that simulation the gate set consisted of the Hadamard and phase gates as well as CNOT gates between any pair of qubits. That is, this simulation uses all-to-all connectivity. However, note that direct RB with this gate set can be applied to a processor with any connectivity, as direct RB does not have to be applied to a processor's native gate set (applying direct RB to a standardized set of n-qubit layers could be useful for comparing different processors). In that case, CNOT gates between distant qubits would be synthesized via SWAP gate chains.

Example gate set 3 [$\mathbb{G}_3(n)$]: An n-qubit gate set constructed from a maximally entangling two-qubit Clifford gate (e.g., CNOT or CPHASE) and the full single-qubit Clifford group.

Direct RB of a gate set with this form is particularly robust and simple to understand from a theoretical perspective (this is the type of gate set that was used in the direct RB experiments of [20]). As with $\mathbb{G}_2(n)$, $\mathbb{G}_3(n)$ specifies a family of gate sets. We now specify a particular convenient gate set within this family (which is entirely specified given a processor's connectivity).

Example gate set 4 [$\mathbb{G}_4(n)$]: An n-qubit gate set consisting of all n-qubit gates composed from applying a layer L_1 and then a layer L_2 where these layers have the following forms: L_1 is a layer containing one of the 24 single-qubit Clifford gates on each qubit, and L_2 is a layer containing non-overlapping CNOT gates on connected pairs of qubits.

A random gate from $\mathbb{G}_4(n)$ locally randomizes the basis of each qubit (and applies some random arrangement of two-qubit gates)—if the marginal distribution over L_1 layers is uniform it

[1] It is only almost any universal gate set rather than every universal gate set as, e.g., it is possible to construct universal gate sets that require arbitrarily long sequences of gates to implement some one-qubit gates—but such gate sets are of no practical relevance.

implements *local* 2-design randomization (composed with a more complex multi-qubit randomization step). A gate set of this form is assumed in parts of Section 4, in order to simplify the theory presented there. Note that direct RB circuits for $\mathbb{G}_4(n)$ have much in common with the random circuits of Google's quantum supremacy experiments [76], which are used in cross-entropy benchmarking [77]. However, these direct RB circuits contain only Clifford gates (making them efficient to simulate), and they contains additional structure (the state preparation and measurement preparation subcircuits).

3.5 Selecting the sampling distribution

There are two main customizable aspects of a direct RB experiment: the gate set ($\mathbb{G}$) and the sampling distribution (Ω). We now discuss the role of Ω, and how to choose it. The direct RB sampling distribution Ω can be any probability distribution over $\mathbb{G}$ that has support on a subset $\mathbb{G}'$ of $\mathbb{G}$ that also satisfies all the requirements for a direct RB gate set (e.g., $\mathbb{G}'$ also generates $\mathbb{C}$). The sampling distribution Ω and the gate set $\mathbb{G}$ control what direct RB measures—direct RB estimates the mean infidelity of an n-qubit gate sampled from Ω. Therefore, the primary consideration when selecting Ω is to choose a distribution that defines an error rate ϵ_Ω of interest. There are infinite valid choices for Ω, and so we do not attempt to discuss all interesting choices for Ω here. One category of sampling distributions we have found useful in experiments is one-parameter families of distributions Ω_ξ in which ξ sets the expected two-qubit gate density of the sampled gate. There are many possible such families of distributions—in Appendix B we briefly describe a family of probability distributions that has been used in direct RB experiments (the "edge grab" sampler, introduced in [78]).

The sampling distribution Ω defines the error rate that direct RB measures, but it also impacts the reliability of direct RB, i.e., whether S_d decays exponentially and how close r_Ω is to ϵ_Ω. So a sampling distribution should be chosen for which direct RB will be reliable. For every sampling distribution satisfying the above requirements, direct RB will be reliable for *sufficiently low error* gates. This is because, for any $(\mathbb{G}, \Omega)$ satisfying the criteria of direct RB, any Pauli error will be randomized by a depth l Ω-random circuit for sufficiently large l (see the theory in Section 4). But this randomization can be very slow, i.e., the required depth l can be very large, and so the error randomization can be much slower than the rate of errors—which, in Section 4, we explain can result in unreliable direct RB (e.g., multi-exponential decays). For example, we could have $\Omega(G) = 1 - \delta$ for one gate G and some $\delta \ll 1$. In this case, a length $l \ll 1/\delta$ sequence of gates sampled from Ω is almost certainly just l repetitions of G—so sequences of this length do not randomize errors at all. The theory presented in Section 4 will help to explain how to choose a sampling distribution that guarantees reliable direct RB.

3.6 The reason for randomized 2-design state preparation and measurement in direct RB circuits

Direct RB is built on the simple idea of directly benchmarking a gate set $\mathbb{G}$ by running varied-depth circuits whose layers are sampled independently from a distribution over $\mathbb{G}$—a class of circuits that have been termed Ω-*distributed random circuits* [51]. However, the depth d direct RB circuits do not just consist of a depth d Ω-distributed random circuits. They surround the core direct RB circuit—the Ω-distributed random circuit—with additional structure (see Fig. 1). We now explain the purpose of this additional structure. The direct RB circuits begin with a randomized state preparation sub-circuit ($C_{\rm sp}$) and end with a measurement preparation sub-circuit ($C_{\rm mp}$). The purpose of $C_{\rm mp}$ is simply to return the qubits to the computational basis, maximizing error visibility and simplifying the data analysis. The subcircuit has the same purpose as the group inversion element in Clifford group RB.

The reasons for beginning a direct RB circuit with a randomized state preparation circuit ($C_{\rm sp}$) are now explained. First, together $C_{\rm sp}$ and $C_{\rm mp}$ implement an approximate (state) 2-design twirl on the error in the core circuit. This is because the state preparation subcircuit generates a sample from a (state) 2-design, which (if implemented perfectly) twirls the overall error channel of the core circuit, so that it behaves as though it is a global depolarizing channel. This will be shown in each of our complementary theories for direct RB (Section 4 and Section 6). An alternative interpretation of $C_{\rm sp}$ is that it seeds the random walk over the group $\mathbb{C}$ induced by a random sequence of group generators (gates from $\mathbb{G}$) with an initial uniformly random group element (that has been compiled into the state preparation for efficiency).

There is also a more practical but mundane reason for including the state preparation subcircuit $C_{\rm sp}$. If $C_{\rm sp}$ is *not* included in a direct RB circuit, the contribution of errors in $C_{\rm mp}$ to a direct RB circuit's total failure probability can be strongly dependent on the depth of the core circuit (d). This would pollute r_Ω, i.e., r_Ω would not accurately quantify the error per layer in the core of the direct RB circuit (ϵ_Ω). To see this, consider direct RB without the randomized state preparation, and assume an efficient compiler for creating $C_{\rm mp}$. Further, assume that $\mathbb{G}$ is small (compared to $\mathbb{C}$), so that the distribution over unitaries produced by a short sequence of gates sampled from Ω must be far from the uniform distribution over $\mathbb{C}$. When d is small, the average depth of $C_{\rm mp}$ will grow approximately linearly with d under many circumstances. This is because $C_{\rm mp}$ is likely to look very much like the depth-d core circuit that it inverts, with the order reversed and individual gates inverted. However, as d gets large, the core circuit converges to a random group element. So the mean depth (and all other properties) of the $C_{\rm mp}$ circuits asymptote to a fixed value that does not depend on d. Critically, this means that the depth of $C_{\rm mp}$ would have a nontrivial dependence on d. This variation will pollute the decay rate (and so r_Ω), and may even cause non-exponential decays—its impact is uncontrolled, as the direct RB protocol is agnostic to how $C_{\rm mp}$ is implemented (just as the Clifford group RB is agnostic about the Clifford compiler). This effect could instead be mitigated by carefully designing the compiler for $C_{\rm mp}$, but this is a complicated undertaking. Instead, the problem is simply solved by the inclusion of $C_{\rm sp}$. With $C_{\rm sp}$ included, the average depth of $C_{\rm mp}$ is independent of d. The only residual d dependence outside of the core circuit is in *correlations* between the state that is prepared by $C_{\rm sp}$ and the state that must be mapped to the computational basis by $C_{\rm mp}$ (they are perfectly correlated at $d = 0$ and uncorrelated as $d \to \infty$). These correlations can have an effect on r_Ω, in principle, but in realistic settings they appear to have no observable effect on r_Ω (see Section 4).

3.7 The reason for conditional compilation: improved scalability

The state preparation and measurement sub-circuits within a direct RB circuit are very similar to the first and last gates in a standard Clifford group RB circuit (see Fig. 1)—and they play essentially the same roles (see above). However, they differ in a practically important way. Both Clifford group RB and direct RB begin with a circuit that implements a group element $F_{\rm sp}$ sampled uniformly from $\mathbb{C}$, but whereas Clifford group RB demands that this circuit implements $U(F_{\rm sp})$ on any input state, direct RB only requires that the circuit generates the same state as $U(F_{\rm sp})$ when applied to $|0\rangle^{\otimes n}$. We call this *unconditional* and *conditional* compilation, respectively. The same distinction holds for the final inversion step, used in direct and Clifford group RB.

We choose to use conditional compilation in direct RB circuits because it substantially increases the number of qubits on which direct RB is feasible—as illustrated by the Clifford and direct RB simulations shown in Fig. 2 (this difference between direct and Clifford group RB is the only reason for the difference in their S_0 values, which is the primary factor setting the number of qubits that it is feasible to benchmark). This is because conditional compilation results in circuits that are shallower and contain fewer two-qubit gates [25]. This is demonstrated in Fig. 3, which compares the mean number of CNOT gates in the circuits generated by a conditional and

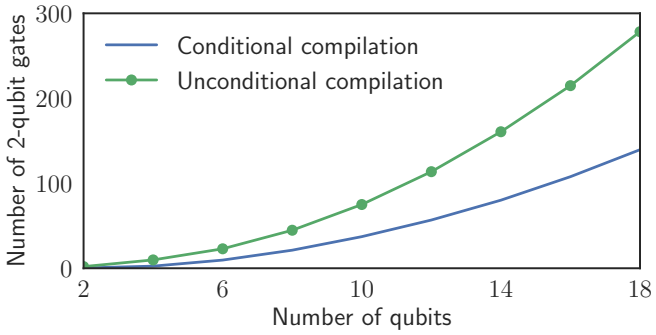

Figure 3: A comparison of the mean number of two-qubit gates in circuits for implementing a uniformly random n-qubit Clifford gate either unconditionally (green connected points) or conditionally on the input of $|0\rangle^{\otimes n}$ (blue line). In the latter case, this is a circuit mapping $|0\rangle^{\otimes n}$ to a uniformly random n-qubit stabilizer state. This demonstrates the significant increase in the number of qubits on which direct RB is feasible—for a fixed two-qubit gate error rate—compared to Clifford group RB. This data was generated using the same open-source compilers that we use for all our direct RB and Clifford group RB simulations [79, 80], and here we compiled into one-qubit gates and CNOT gates between any pair of qubits (i.e., all-to-all connectivity).

unconditional n-qubit Clifford compiler. To generate this plot we used open-source compilation algorithms [79, 80] (that we also used for all the simulations herein). These algorithms are unlikely to generate circuits with minimal two-qubit gate counts (see [25, 26, 27] for work on optimal Clifford compilation) although we have found that they perform reasonably well.[2]

Note that the error rate measured by direct RB is (approximately) independent of the details of the compiler used to generate these states. The property of the compilation algorithm that is of importance to direct RB is the efficiency with which it generates these states—the higher the fidelity of these states the more qubits direct RB will be feasible on. This is convenient, as algorithms for generating many-qubit states/unitaries are typically "black-box" and the properties of the circuits they generate cannot be easily controlled. In contrast, with Clifford group RB the compiler entirely defines the physical meaning of the RB error rate—a Clifford group RB error rate cannot be used to quantify *native gate* performance without a detailed understanding of the compiled circuits.

3.8 The reason for randomizing the success outcome

Direct RB specifically allows for uniform randomization of the success bit string (this is also possible with Clifford group RB [81], but it is not standard practice). This bit string randomization is not essential, but in our view it is preferable. This is because it guarantees that the $d \to \infty$ asymptote of the average RB success probability S_d is $1/2^n$ (for all Markovian errors), so we can fix $A = 1/2^n$ in $S_d = A + Bp^d$. As Harper *et al.* [81] discuss the motivation and statistical impact of this bit string randomization (in the context of Clifford group RB), we do not do so further here.

[2]The algorithm we use is based on Gaussian elimination; the compiled circuits contain $O(n^2)$ two-qubit gates for both the state and unitary compilations. This scaling is not optimal: there are algorithms that generate circuits containing $O(n^2/\log(n))$ two-qubit gates [25]. But in the $n = O(10)$ regime we have found that our compilations contain many fewer two-qubit gates than the $O(n^2/\log(n))$ algorithms of [25]. We have not compared our compilation algorithms to more recent work on Clifford compilation [27].

3.9 The error rate convention: choosing the decay rate scaling factor

For both direct and Clifford group RB, data are analyzed by fitting the average success probabilities to $S_d = A + Bp^d$ and then mapping p to an error rate r (or a fidelity $1 - r$). In Clifford group RB (and other RB protocols), p is conventionally mapped to an error rate defined by $r' = (2^n - 1)(1 - p)/2^n$ [7, 8]. This is not the definition that we use. As specified in Eq. (15), we define the direct RB error rate to be $r = (4^n - 1)(1-p)/4^n$. For $n \gg 1$ the difference between r and r' is negligible, but it is substantial for $n \sim 1$. Our decision to use Eq. (15) to define r is not specific to direct RB. The RB error rate r' corresponds to the gate set's mean *average gate infidelity*, whereas r corresponds to the mean *entanglement infidelity*. We choose to use a definition for the RB error rate that correspond to entanglement infidelity because of entanglement infidelities convenient properties. For example, for Pauli stochastic errors the entanglement infidelity corresponds to the probability of any Pauli error occurring (see Section 2) and the entanglement fidelity of multiple gates used in parallel is equal to the product of their entanglement fidelities (assuming no additional errors when gates are parallelized)—see Eq. (17). Due to these convenient properties, this choice has also been made with other scalable benchmarking techniques (e.g., cycle benchmarking [82]).

3.10 Comparison to other methods for native gate randomized benchmarking

Direct RB is not the first or only proposal for RB directly on a gate set that generates a group. Below we explain the relationship between RB and each of these protocols:

- The method of Knill *et al.* [6] benchmarks a set of gates that generate the one-qubit Clifford group, and this has been widely used as an alternative to Clifford group RB (e.g., see the references within [83]). That method consists of uniform sampling over a specific set of one-qubit Clifford group generators. So it is a particular example of direct RB (as pointed out by Boone *et al.* [83]) *except* that it does not include the initial stabilizer state preparation step (which is of little importance in the single-qubit setting). So our theory for direct RB is further evidence that the protocol of [6] is just as reliable and arguably as well-motivated as Clifford group RB, although it measures a different error rate.

- The extensions of the method of Knill *et al.* up to three qubits [50] also fit within the framework of direct RB except that, again, in that method there is no stabilizer state initialization.

- Independently of the development of direct RB, França and Hashagen proposed "generator RB" [48], which is also direct RB without the stabilizer state preparation step and without user-configurable sampling (but note that some of the theory within [48] also applies to direct RB).

- Since the development of direct RB, a protocol called mirror RB [43, 51] has been introduced that adapts direct RB to improve its scalability. Mirror RB methods replace the stabilizer state preparation and measurement sub-circuits from direct RB with a mirror circuit reflection structure, meaning that they contain no large subroutines. Mirror RB techniques are more scalable than direct RB, and they are designed to measure the same error rate as direct RB (ϵ_Ω). However, the error rate measured by existing mirror RB methods is a less reliable estimate of ϵ_Ω (it is typically a slight underestimate—see [43, 51] for further details), so mirror RB is not a strict improvement on direct RB. Much of the theory presented in this paper is also applicable, or adaptable, to mirror RB.

- After the initial version of this paper appeared, a protocol called binary RB [52] was introduced that (like mirror RB) adapts direct RB to improve its scalability. Binary RB removes the stabilizer state preparation and measurement subcircuits from direct RB, replacing each

with a single layer of single-qubit gates, and changes the data analysis using the ideas from direct fidelity estimation [84]. This makes binary RB more scalable than direct RB, and it is designed to measure the same error rate as direct RB (ϵ_Ω). Much of the theory presented here is also applicable (or adaptable) to binary RB.

4 Understanding direct RB using error scrambling

In this section we present a simple approximate theory of direct RB. The theory in this section is designed to (1) highlight the key properties of a gate set, and a sampling distribution, that guarantee reliable direct RB, and (2) to explain why direct RB works when these conditions are satisfied. In this theory we make a range of simplifying assumptions that allow us to present simple conditions under which direct RB reliably estimates the average infidelity of the benchmarked gates. The main assumptions we make in this section are as follow (see also Table 1):

1. The gate set contains only Clifford gates (and generates the Clifford group).
2. All errors are stochastic Pauli errors.
3. Many ($n \gg 1$) qubits are involved. Around $n \sim 5$ suffices.
4. Each gate's error rate is low ($\epsilon \ll 1$).

The first two assumptions are foundational for our arguments (although it is plausible that a conceptually similar, but likely more mathematically complex, theory could be created for general gates and general stochastic errors). This is because our arguments are based on how Pauli errors propagate through layers of Clifford gates. The third assumption—many qubits—is not necessary, but simplifies notation. This assumption enables us to drop $O(1/4^n)$ terms that appear because two random non-identity Pauli operators (errors) compose to the identity with probability $1/4^n - 1$. This effect is not negligible when $n \sim 1$. It can be accounted for in all of the following theory, but it causes notational complications that, in our opinion, obfuscate the key intuitions. As our theory in Section 6 is applicable to the few-qubit setting, we therefore ignore the $n \sim 1$ setting here. The final assumption is simply that the rate of error in a single layer of gates is small, which we require because otherwise we cannot guarantee any error scrambling occurring between two errors within a circuit. As we discuss in more detail later in this section, we expect that this assumption can be relaxed to the condition that the *marginal* rate of errors on each qubit is small (which is a physically well-justified assumption for any n), but we do not rigorously show this herein.

4.1 Clifford gates and stochastic Pauli errors

We now introduce our notation for this section, expand on some of our assumptions, and summarize the results of this section. In this section we are studying direct RB of an n-qubit gate set $\mathbb{G}$ that generates the Clifford group, with a sampling distribution Ω that has support on all of $\mathbb{G}$. Each $G \in \mathbb{G}$ is modeled by the ideal unitary $U(G)$ followed by a G-dependent Pauli stochastic channel (assumption 2). This Pauli stochastic channel is described by 4^n rates $\{\epsilon_{G,P}\}_{P\in\mathbb{P}}$ where $\epsilon_{G,P}$ is the probability that $U(G)$ is followed by the Pauli P and $\mathbb{P}$ is the n-qubit Pauli group. So

$$\sum_{P\in\mathbb{P}} \epsilon_{G,P} = 1, \tag{18}$$

and $\epsilon_{G,I}$ is the probability of no error where I is the identity Pauli operator. The probability of any Pauli error is

$$\sum_{P\in\mathbb{P}, P\neq I} \epsilon_{G,P} = \epsilon_G, \tag{19}$$

where ϵ_G is the entanglement infidelity of G. In this section, we will show that for a broad range of $\mathbb{G}$, Ω, and stochastic Pauli error models (defined by $\{\epsilon_{G,P}\}_{G\in\mathbb{G},P\in\mathbb{P}}$), direct RB is reliable. That is, the direct RB average success probability (S_d) decays exponentially ($S_d \approx A + Bp^d$) and the direct RB error rate (r_Ω) is approximately equal to the mean entanglement infidelity of a gate sampled from Ω ($r_\Omega \approx \epsilon_\Omega$).

Under the assumptions introduced above, a benchmark depth d direct RB circuit consists of:

1. A circuit preparing a uniformly random n-qubit stabilizer state ψ.

2. A circuit $C = G_d \cdots G_1$ of d gates G_i sampled from Ω, which is called the direct RB circuit's "core".

3. A circuit and measurement that simulates the POVM $M_\varphi = \{|\varphi\rangle\langle\varphi|, 1 - |\varphi\rangle\langle\varphi|\}$, where φ is the stabilizer state
$$|\varphi\rangle = U(C)|\psi\rangle. \tag{20}$$
We refer to this as a stabilizer state measurement.

Until stated later in this section, we will assume that the stabilizer state preparation and measurement (SSPAM) are perfect, i.e., we exactly prepare ψ and exactly measure M_φ. This is not a realistic assumption, but, as we argue below, the primary effect of SSPAM error is on the initial value of the success probability decay (i.e., $A + B$ in $S_d = A + Bp^d$). Therefore, SSPAM error has negligible impact on r_Ω, and only impacts the number of qubits on which direct RB is feasible.

4.2 Modelling direct RB circuits with a stochastic unravelling

Throughout this section we use a *stochastic unraveling* of the stochastic Pauli error model, and we treat a direct RB circuit, and the gates it contains, as random variables. This is illustrated in Fig. 4 (a) [for $d = 3$]. In particular, a random depth d direct RB circuit's output "success" bit (B_d, with $B_d \in \{0, 1\}$) is given by
$$B_d = |\langle\varphi|\tilde{U}(C)|\psi\rangle|^2, \tag{21}$$
where
$$\tilde{U}(C) = P_d U(G_d) \cdots P_2 U(G_2) P_1 U(G_1). \tag{22}$$
Here the $U(G_i)$ are unitary-valued random variables, that are mutually independent and Ω-distributed, the P_i are Pauli-operator-valued random variables where P_i is equal to the Pauli operator P with probability $\epsilon_{G_i,P}$ (so P_i depends on G_i), and ψ is a stabilizer-state valued random variable that is independent and uniformly distributed (and $\langle\varphi| = \langle\psi|U^{-1}(C)$). The P_i correspond to the potential errors in the direct RB circuit. The average success probability of a depth d direct RB circuit (S_d) is the expected value of B_d, i.e.,
$$S_d = \mathbb{E}[B_d]. \tag{23}$$

In our theory below, we will show how S_d can be calculated by considering the probabilities of different mutually exclusive "patterns" of Pauli errors within Eq. (22).

4.3 Three routes to success: no errors, multiple errors cancel, or the measurement misses an error

The first step in our theory is to show that the direct RB average success probability (S_d) can be decomposed into the sum of three contributions [see Fig. 4 (b)]: circuit execution instances in which (1) no Pauli errors occurred, (2) at least two Pauli errors occur but they cancel out, and (3)

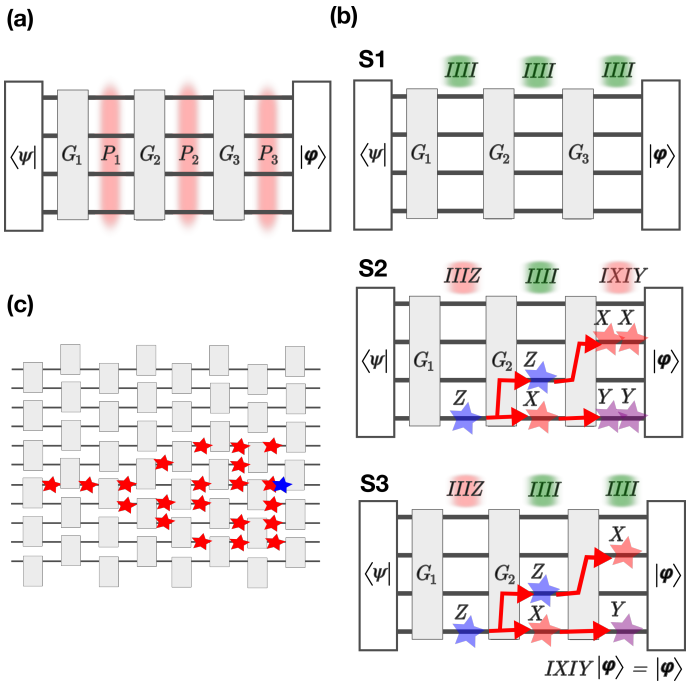

Figure 4: Our theory for direct RB with stochastic Pauli errors uses (a) a stochastic unravelling of a Pauli stochastic error model [see Sections 4.1–4.2], illustrated here for a depth $d = 3$ circuit. To prove that direct RB is reliable we (b) consider the probabilities of three mutually exclusive events [S1-S3; see Section 4.3] within that unravelling that each contribute to the probability that a depth d direct RB circuit returns the "success" bit string (S_d). Our theory shows that if the probability of two errors cancelling within a direct RB circuit (the event S2) is negligible then $S_d = A + B(1 - \epsilon_\Omega)^d$ [see Sections 4.4–4.6]. (c) The probability of two or more errors cancelling is negligible, because even a single-qubit Pauli error (leftmost red star) is almost certainly spread onto many qubits by a small number of random circuit layers (red stars illustrate a possible error spreading trajectory), so that it cannot cancel with a few-qubit error later in the circuit (blue star) [see Section 4.7].

Pauli errors occur and they do not cancel, but they are unobserved due to the particular SSPAM in that circuit. We begin by noting that Eq. (21) can be rewritten as

$$B_d = |\langle\varphi|\tilde{U}(C)|\psi\rangle|^2 = |\langle\psi|U^\dagger(C)\tilde{U}(C)|\psi\rangle|^2. \tag{24}$$

Now, because the gates are all Clifford operators, we have that

$$\tilde{U}(C) = PU(C), \tag{25}$$

where P, which is a Pauli-operator-valued random variable, is calculated by commuting each of the P_i past the subsequent gates, and then multiplying them together. So

$$B_d = |\langle\psi|U(C)^\dagger PU(C)|\psi\rangle|^2 = |\langle\psi|\tilde{P}|\psi\rangle|^2, \tag{26}$$

where $\tilde{P}$ is the Pauli-operator-valued random variable given by

$$\tilde{P} = U(C)^\dagger PU(C). \tag{27}$$

Equation (26) implies that a direct RB circuit's outcome is the "success" bit string (i.e., $B_d = 1$ rather than $B_d = 0$) if and only if one of the following *mutually exclusive* events occurs:

S1. All $P_i = I$, meaning that no errors occurred in C. This implies that $P = I$, so $\tilde{P} = I$. [See Fig. 4 (b) S1].

S2. Two or more of the P_i operators are errors, i.e., they are not the identity, but we still have that $P = I$ up to a phase, so $\tilde{P} = I$ up to a phase. This means that two or more errors occurred, but—after they are propagated past the circuit layers separating them—they compose to the identity. [See Fig. 4 (b) S2].

S3. One or more of the P_i operators are errors *and* they do not cancel ($P \neq I$), but they are nonetheless unobserved by the stabilizer measurement. This means that $|\psi\rangle$ is an eigenstate of $\tilde{P}$ (equivalently $|\varphi\rangle$ is an eigenstate of P). This means that $\tilde{P}$ is a stabilizer or anti-stabilizer of the prepared stabilizer state. [See Fig. 4 (b) S3].

The benchmark depth d average success probability ($S_d = \mathbb{E}[B_d]$) is the probability that any one of these mutually exclusive events (S1, S2, and S3) occurs in a random depth d direct RB circuit. That is, we can write

$$S_d = \mathrm{Pr}_d(\mathrm{S1}) + \mathrm{Pr}_d(\mathrm{S2}) + \mathrm{Pr}_d(\mathrm{S3}), \tag{28}$$

where $\mathrm{Pr}_d(E)$ denotes the probability of the event E occurring in a benchmark depth d direct RB circuit. Because S1, S2, and S3 are exclusive events, by applying basic probability theory we can rewrite S_d as

$$S_d = s_1 + (1 - s_1)s_2 + (1 - s_1)(1 - s_2)s_3, \tag{29}$$

where s_1 is the probability that S1 occurs, s_2 is the probability that S2 occurs given that S1 does not, and s_3 is the probability that S3 occurs given that S1 and S2 do not. That is,

$$s_1 = \mathrm{Pr}_d(\mathrm{S1}), \tag{30}$$
$$s_2 = \mathrm{Pr}_d(\mathrm{S2} \mid \bar{\mathrm{S1}}), \tag{31}$$
$$s_3 = \mathrm{Pr}_d(\mathrm{S3} \mid \bar{\mathrm{S1}} \cap \bar{\mathrm{S2}}). \tag{32}$$

In the next two subsection we show that s_1 and s_3 have very simple forms, before then turning to s_2.

4.4 A formula for the probability of no errors

We have shown that the direct RB decay (S_d) can be expressed in terms of three quantities—s_1, s_2, and s_3 [Eq. (29)]. We now derive a formula for s_1, defined in Eq. (30). This is the probability that no errors occur in a depth d Ω-distributed random circuit [see Fig. 4 (b) S1]. A depth d Ω-distributed random circuit consists of d gates from $\mathbb{G}$ that are sampled independently from Ω. Because ϵ_Ω is the probability that a gate sampled from Ω experiences a Pauli error, $(1 - \epsilon_\Omega)$ is the probability that no error occurs for a gate sampled from Ω. As the d gates in our circuit are independently sampled, we therefore simply have

$$s_1 = (1 - \epsilon_\Omega)^d. \tag{33}$$

4.5 A formula for the probability that an error is not observed

We now derive a formula for the probability that an error in a direct RB circuit is not observed by the measurement (i.e., s_3), which is one of the three quantities in our formula for S_d in Eq. (29). Defined in Eq. (32), s_3 is the probability that any errors in the core of a benchmark depth d direct RB circuit do not cause the circuit to return the wrong bit string given that (i) at least

one error occurs, and (ii) the errors do not cancel [see Fig. 4 (b) S3]. Together these conditions imply that $\tilde{P}$ in Eq. (26) is not the identity. So s_3 is the probability that $|\langle\psi|\tilde{P}|\psi\rangle|^2 = 1$ given that $\tilde{P} \neq I$. The state ψ is a uniformly random stabilizer state (that is independent of $\tilde{P}$), so it is an eigenstate of *any* fixed Pauli error (i.e., a non-identity Pauli operator) with probability $(2^n-1)/(4^n-1) = 1/(2^n+1)$. This is because any stabilizer state ψ is an eigenstate of 2^n-1 non-identity Pauli operators, and there are 4^n-1 non-identity Pauli operators [25]. So

$$s_3 = \frac{1}{2^n+1} = \frac{1}{2^n} + O(1/4^n). \tag{34}$$

This simple d-independent formula for s_3 is a consequence of the random state preparation step in direct RB—it guarantees that s_3 depends only on whether $\tilde{P}$ is an error, not on its unknown (and d-dependent) distribution over the 4^n-1 Pauli errors.

4.6 Direct RB is reliable if the probability of error cancellation is negligible

We now explain why direct RB is reliable whenever the probability of multiple errors canceling in a direct RB circuit is negligible. Above, we derived simple and d-independent equations for s_1 [Eq. (33)] and s_3 [Eq. (34)]. Substituting these equations into Eq. (29) we obtain

$$S_d = \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right)(1-\epsilon_\Omega)^d(1-s_2) + s_2 + O(1/4^n). \tag{35}$$

This equation for S_d depends on n, ϵ_Ω, and s_2, where s_2 is the probability that all the errors within a direct RB circuit's core cancel out given that at least one error occurs. If we assume that $s_2 \approx 0$ [and that $n \gg 1$, so that $1/4^n \approx 0$], then Eq. (35) becomes

$$S_d \approx \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right)(1-\epsilon_\Omega)^d, \tag{36}$$

i.e., $S_d \approx A + Bp^d$ with $A = 1/2^n$, $B = 1 - 1/2^n$, and $p = 1 - \epsilon_\Omega$, so

$$r_\Omega \approx \epsilon_\Omega. \tag{37}$$

This implies that, when the s_2 is small, direct RB is guaranteed to be reliable. The remainder of our theory consisting of showing that s_2 is negligible under very broad conditions.

4.7 The probability of error cancellation in direct RB circuits is negligible

We now explain why the probability of errors cancelling in direct RB circuits is typically negligible (i.e., $s_2 \approx 0$). The core of the argument is:

(i) random circuit layers quickly turn an error that occurs on few qubits into an error impacting many qubits [61, 62, 63], and

(ii) the probability that a many-qubit error cancels with another error is negligible, so

(iii) if errors are delocalized faster than they occur then $s_2 \approx 0$, resulting in reliable direct RB.

Therefore, direct RB will be reliable as long as $\epsilon_\Omega < 1/k_{\text{delocal}}$ where k_{delocal} is the number of layers required to delocalize any error and ϵ_Ω is the infidelity of a gate sampled from Ω. Importantly, k_{delocal} is a small constant—it is independent of n. This argument is illustrated in Fig. 4 (c), and we expand upon it below.

Our argument will use some simplifying assumptions, one of which uses the concept of a Pauli P's *weight* $W(P)$. The weight $W(P)$ of the Pauli P is the number of qubits on which P acts non-trivially, i.e., $W(P) = w$ if P has a non-identity action (X, Y or Z) on w qubits and it is an identity on the remaining $n-w$ qubits. We assume:

1. An even number of qubits n with an α-regular connectivity graph and an α-edge-coloring, e.g., a ring ($\alpha = 2$), torus ($\alpha = 4$), or fully connected graph ($\alpha = n - 1$).

2. An n-qubit gate set $\mathbb{G} = \{G\}$ consisting of all n-qubit gates of the form $G = L_2 L_1$ where:

   (a) L_1 is a layer containing one of the 24 single-qubit Clifford gates on each qubit, and

   (b) L_2 is a layer of CNOT gates that contains all $n/2$ CNOT gates corresponding to one of the α edge colors on the connectivity graph (so there are α different two-qubit gate layers).

3. Ω is the uniform distribution over $\mathbb{G}$.

4. The probability of high-weight errors is negligible, i.e., for every gate G, $\sum_{P \in \mathbb{P}_4} \epsilon_{G,P} \approx 0$ where P_w is the set of all Pauli operators of weight w or greater.

5. $n \gg 1$, as assumed throughout this section.

The first three of these assumptions constitute a type of spatial homogeneity: these assumptions imply that when a Pauli error occurs on a qubit q, Ω-distributed random layers cause it to experience the same randomization process independent of the basis of the error (X, Y or Z) and on which qubit it occurs. This simplifies the following argument, but it is not essential to it.[3] The fourth assumption is also not essential, but it is physically reasonable and it also simplifies the argument. The designation of weight-4 and above errors as high-weight, and weight-3 and below errors as low-weight is somewhat arbitrary—any other weight $w \ll n$ can be used to define "low-weight" in the following derivation.

We aim to show that s_2 is negligible, where s_2 [defined in Eq. (31)] is the probability that (i) two or more errors occur in $\tilde{U}(C)$ [Eq. (22)] and (ii) these errors cancel, so that $\tilde{U}(C) = U(C)$. We consider the case of (at least) two errors occurring in Eq. (22), by conditioning on errors occurring after layers i and $i+k+1$, in Eq. (22), for some $i, k \geq 1$. This means that we are conditioning on the Pauli P_i and P_{i+k+1} being not equal to the identity—so they are now random variables that are distributed over the $4^n - 1$ possible Pauli errors—and we will denote these conditional random variables by P and Q', respectively. Now, because Q' is a random variable that models an error occurring on gate G_{i+k+1} it is not independent of $U(G_{i+k+1})$. So we commute Q' in front of $U(G_{i+k+1})$, by defining $Q = U(G_{i+k+1})^{-1} Q' U(G_{i+k+1})$ [Q is essentially just a pre-gate Pauli error, rather than a post-gate Pauli error]. Therefore P and Q are independent Pauli-operator-valued random variables with k independent random unitaries $U_{i+k:i} \equiv U(G_{i+k}) \cdots U(G_i)$ separating them [see the illustration in Fig. 4 (c)].

We now show that the probability that P and Q cancel is negligible—i.e., the probability of $Q U_{i+k:i} P = U_{i+k:i}$ (up to a phase) is negligible—when k is sufficiently large. This is the condition

$$\Pr[P^{(k)} = i^a Q] < \delta_{\text{cancel}}, \tag{38}$$

for some small $\delta_{\text{cancel}} \geq 0$ (and where i^a is any phase), where

$$P^{(k)} = U_{i+k:i} P U_{i+k:i}^{-1}. \tag{39}$$

Now, by assumption [see (4) above] the probability that Q is a high-weight error (weight 4 or higher) is negligible, and so

$$\Pr[P^{(k)} = i^a Q] \leq \Pr[W(P^{(k)}) = W(Q)] \leq \kappa(3, k) \tag{40}$$

[3] The following derivation can be reformulated in terms of the Pauli error that is randomized slowest by Ω distributed layers.

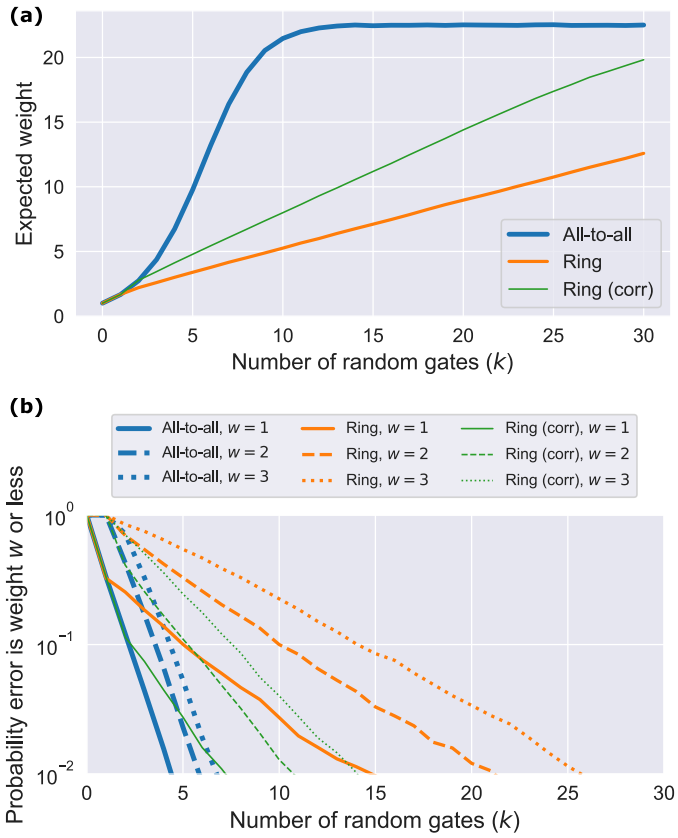

Figure 5: In the main text we prove that direct RB is guaranteed to be reliable if direct RB's random gates convert a weight-1 Pauli error into a high-weight Pauli error (a weight-4 or above error) before there is a high probability that another error occurs. To show that weight-1 errors are quickly converted into high-weight errors by random gates from practically relevant gate sets we simulated this randomization process. (a) The expected weight of a weight-1 Pauli error after it has been propagated through k n-qubit gates $G = L_2 L_1$, where L_1 is a layer of uniformly random single-qubit Clifford gates, and L_2 is a random layer containing $n/2$ CNOT gates between pairs of connected qubits (see Section 4.7). We show the expected weighted versus k for all-to-all connectivity (blue), ring connectivity (orange), and ring connectivity with *correlated* sampling of the CNOT layers (green) in which the L_2 layers alternate between coupling a qubit with its neighbour to its left or right (this speeds up error spreading). These results do not depend on the location of the initial Pauli error, because we assume a form of spatial homegeniety (an α-regular connectivity graph that is α-edge-colorable). (b) The probability that a weight-1 Pauli error is still a weight-w or less error after it has been propagated through k n-qubit gates, for $w = 1, 2, 3$. As expected (see main text), we find that these probabilities drop off quickly with k. These results are for $n = 30$ qubits [the probabilities in (b) upper-bound their values for $n > 30$].

where $\kappa(w, k)$ is the probability that $P^{(k)}$ is a weight-w or fewer error:

$$\kappa(w, k) \equiv \sum_{w'=1}^{w} \Pr[W(P^{(k)}) = w']. \tag{41}$$

So, the probability that P and Q cancel is bounded by the probability that $P^{(k)}$ is a weight-3 or fewer error.

The probability that $P^{(k)}$ is a weight-3 or fewer error decreases rapidly with increasing k, independent of P's distribution over Pauli errors. The probability distribution for P that minimizes

the expected weight of $P^{(k)}$ is a distribution that only has support on weight-1 errors. But a weight-1 Pauli operator rapidly increases in weight as it is commuted through k Ω-distributed random gates, as illustrated in Fig. 4 (c). To see why this is, consider the effect of commuting some Pauli operator P_{in} through a random gate $G = L_2 L_1$ [L_1 and L_2 are as defined in assumption (2), above].

- *Propagation through* L_1: The Pauli operator P_{in} is first commuted through a layer of independent and uniformly random single-qubit Clifford gates (L_1). This has a local homogenising effect on P_{in}. Specifically, if we let $E(P_{\text{in}})$ denote the set of qubits that P_{in} acts non-trivially on, then this layer transforms P_{in} into a *uniformly random* non-identity Pauli on that set of qubits. Therefore, $P_{\text{in}} \to P_{\text{int}}$ where P_{int} is uniformly distributed over the 3^w different Pauli errors on $E(P_{\text{in}})$ where $w = W(P_{\text{in}})$.

- *Propagation through* L_2: The randomized Pauli operator P_{int} is then propagated through a random layer of CNOT gates (L_2). In expectation, this will typically (but not always) spread the error to more qubits as long as $W(P_{\text{int}}) < 3n/4$.

  - If none of the qubits in $E(P_{\text{int}})$ are connected to each other, L_2 cannot couple pairs of qubits within $E(P_{\text{int}})$. In this case, L_2 contains CNOT gates that couple each qubit within $E(P_{\text{in}})$ with a qubit outside of $E(P_{\text{in}})$, spreading the error onto that other qubit with a probability of $2/3$. This is because a CNOT gate maps 4 of the 6 weight-1 Pauli operators to weight-2 Pauli operators. Therefore, the effect of L_2 is to spread the error to additional qubits: it maps $P_{\text{int}} \to P_{\text{out}}$ where the expected weight of P_{out} is

$$\mathbb{E}[W(P_{\text{out}})] = W[P_{\text{in}}] + 2W[P_{\text{in}}]/3 = 5W[P_{\text{in}}]/3. \tag{42}$$

  - If some of the qubits in $E(P_{\text{int}})$ are connected to each other, there are two effects that compete to decrease or increase the weight of P_{int}. Those CNOT gates in L_2 that interact two qubits within $E(P_{\text{int}})$ will correct (i.e., remove) the error on one of the qubits with probability $4/9$. This is because a CNOT gate maps 4 of the 9 weight-2 Pauli operators to weight-1 Pauli operators. In contrast, those CNOT gates in L_2 that interact a qubit from $E(P_{\text{int}})$ with one outside of $E(P_{\text{int}})$ spread an error to that other qubit with probability $2/3$.

The distribution of $P^{(k)}$ can be calculated by k applications of the above randomization process to P. The Pauli-operator-valued random variable P is (by assumption) distributed over only low-weight errors, but the process of propagating P through j gates ($G = L_2 L_1$) causes the expected weight of the error $P^{(j)}$ to quickly increase with j. The rate of increase depends on the probability at step j that a CNOT in L_2 couples two qubits within $E(P^{(j)})$ instead of coupling one qubit within $E(P^{(j)})$ with one outside of $E(P^{(j)})$. This probability depends on the qubits' connectivity, as well as the initial distribution of P. The result is that $\mathbb{E}[W(P^{(j)})]$ initially grows linearly with j if $\alpha = 2$ (ring connectivity) and exponentially if $\alpha = n-1$ (all-to-all connectivity) [61, 62, 63] [see Fig. 5 (a), and note that in all cases $\mathbb{E}[W(P^{(j)})] \to 3n/4$ as $j \to \infty$]. Most importantly for our purposes, for *any* connectivity α the number of layers k required to satisfy $\kappa(3,k) \approx 0$ is a *constant*, i.e., it is independent of n. [see Fig. 5 (b) for $\kappa(3,k)$ when $n = 30$—this figure is discussed further below].

We are now ready to state a sufficient condition for reliable direct RB. A condition that is independent of the underlying Pauli error model is preferable (as this is unknown in practice), so we define:

$$K(w,k) = \max_P \kappa(w,k) \equiv \max_P \left\{ \sum_{w'=1}^{w} \Pr[W(P^{(k)}) = w'] \right\}. \tag{43}$$

This is $\kappa(w,k)$ maximized over all possible distributions for P—the worst-case is when P has support only on weight-1 errors. $K(w,k)$ quantifies the probability that a Pauli-operator-valued random variable P is a weight-w or less error after it is propagated through k random layers, maximized over all possible distributions for P. We now define k_{delocal} to be the smallest k such that $K(3,k) < \delta_{\mathrm{delocal}}$ where $\delta_{\mathrm{delocal}} > 0$ is some small constant. The above argument has shown that:

(i) two or more errors that are separated by at least k_{delocal} layers have negligible probability to cancel, and

(ii) k_{delocal} is independent of n.

Direct RB will therefore be reliable as long as the probability that two or more errors occur within k_{delocal} gates of each other is negligible. We can guarantee this with the following condition:

$$\epsilon_\Omega < 1/k_{\mathrm{delocal}}. \tag{44}$$

This is a simple and easy-to-verify condition that is sufficient for reliable direct RB. Note that it is easy-to-verify as (i) ϵ_Ω can be estimated by direct RB,[4] and (ii) k_{delocal} can be efficiently calculated using simulations, for a particular Clifford gate set $\mathbb{G}$ and sampling distribution Ω.

Equation (44) shows that direct RB will be reliable whenever ϵ_Ω is smaller than some constant ($1/k_{\mathrm{delocal}}$). However, ϵ_Ω will typically grow with n. Therefore, Eq. (44) will not be satisfied for all n for any realistic n-parameterized error models (e.g., constant single- and two-qubit gate error rates). So it is interesting to consider whether there is a weaker condition that guarantees reliable direct RB.[5] Two errors can cancel only if, when propagated past the intervening layers, they occur on the same set of qubits (as implied by the illustration of Fig. 4 (c)). So, the probability of error cancellation will be minimal even if two errors often occur within k_{delocal} layers of each other as long as those errors almost certainly occur on distant qubits. This suggests that direct RB will be reliable if $\epsilon_{\Omega,\mathrm{perQ}} < 1/k_{\mathrm{delocal}}$ where $\epsilon_{\Omega,\mathrm{perQ}}$ is some quantification of the error rate per-qubit. This can be formalized if we assume a local Pauli error model, i.e., every gate's error channel is a tensor products of local Pauli stochastic channels. Then our argument (above) implies that direct RB will be reliable if

$$\epsilon_{\Omega,\mathrm{perQ,max}} < 1/k_{\mathrm{delocal}}, \tag{45}$$

where $\epsilon_{\Omega,\mathrm{perQ,max}}$ is the maximum of the average infidelities of the n qubits. Note that, in this error model, $\epsilon_{\Omega,\mathrm{perQ,max}}$ is well-defined because the tensor-product structure of the error maps means that each qubit has a well-defined infidelity per n-qubit gate.

Our sufficient condition for reliable direct RB [Eq. (44)] depends on the number of random layers needed to spread a low-weight error onto approximately 4 or more qubits (k_{delocal}). We have argued that k_{delocal} is independent of n (for $n \gg 1$) and that it decreases with increasing qubit connectivity (α). To quantify k_{delocal} and verify its dependence on α, we used simulations to calculate the probability $K(w,k)$ that a weight-1 error is a weight w or less error after it has been commuted past k random gates. Figure 5 (b) shows the results, for $w = 1,2,3$ and two extremal connectivities: all-to-all (blue) and ring (orange) connectivity [we also show $K(w,k)$ for a ring connectivity with *correlated* sampling of L_2 layers, so that the L_2 layers alternate between coupling a qubit with its neighbour on its left or right, which significantly speeds up error propagation]. We find that $K(w,k)$ for $w = 1,2,3$ decreases rapidly for both linear and all-to-all

[4] Even when direct RB is not reliable, it will provide a reasonable order-of-magnitude estimate for ϵ_Ω. Alternatively, an approximate value for ϵ_Ω may already be known.

[5] Even though direct RB will not be feasible when $n \gg 1$ *and* ϵ_Ω is large because the SSPAM subcircuits of direct RB circuit will not be implementable with significantly non-zero fidelity.

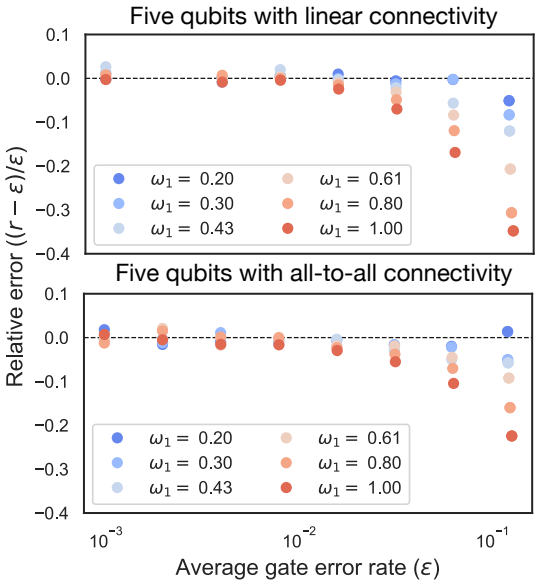

Figure 6: Simulations investigating the reliability of direct RB in the slow scrambling regime. The discrepancy between the estimated direct RB error rate (r) and the actual average gate error rate (ϵ) in simulations, as quantified by the relative error $\delta = (r-\epsilon)/\epsilon$. The simulations are of a 5-qubit device with linear connectivity (upper plot) and all-to-all connectivity (lower plot). By varying the average gate error (ϵ) and the in-homogeneity of the error rates across the qubits (related to ω_1) we see that direct RB is reliable under a wide range of different connectivities and stochastic error models. Details of the simulation, including definitions of ω_1 and ϵ are given in the main text.

connectivity. These results are for $n = 30$ qubits, but note that there is only a weak n dependence and these $K(w, k)$ upper-bound the values of $K(w, k)$ for $n > 30$ qubits.

4.8 Numerical simulations demonstrating negligible error cancellation

Our theory implies that direct RB is reliable whenever the probability of error cancellation is negligible, and that this error cancellation probability is negligible under broad conditions. It also implies that the error cancellation probability decreases as

(1) the error rates of the gates decreases,

(2) the homogeneity of the error probabilities increases (reduced bias), and

(3) the speed of error delocalization and scrambling increases.

We investigated these claims using numerical simulations of direct RB. Our simulations are of direct RB on five qubits with a gate set containing gates constructed from parallel applications of (1) CNOT gates between connected qubits, and (2) all 24 single qubit Clifford gates.

A processor's connectivity strongly effects the maximum speed with which errors can be spread. With all-to-all connectivity, a single random layer can spread a localized error so that, for every qubit, there is an $O(1/n)$ probability that the error has been been spread to that qubit. In contrast, with linear connectivity, a single random layer can only spread an error to one of at most two adjacent qubits—meaning that there is a significant probability that an error that is delocalized is then relocalized by a subsequent random layer. So, to investigate effect (3), above, we simulated direct RB for two extremal processor connectivities: all-to-all and linear connectivity. We investigate effects (1) and (2), above, by varying the strength of the errors and by interpolating

from (i) equal error probabilities on all qubits, and (ii) errors occurring only on a single qubit. In the simulated error model, qubit i (with $i = 1, 2, \ldots, 5$) had an error rate of $\epsilon_i = \tilde{\epsilon}\omega_i$ for weights given by

$$\omega_i = \frac{\alpha^i}{\sum_{j=1}^{5} \alpha^j}, \tag{46}$$

for some α. On each qubit, each gate experiences the same error map, and each Pauli occurs with equal probability, i.e., local depolarization. The value of α controls the homogeneity of the error rates on the different qubits. The value of $\tilde{\epsilon}$ sets the error rate:

$$\epsilon_\Omega = 1 - \prod_i (1 - \tilde{\epsilon}\omega_i) \approx \tilde{\epsilon}. \tag{47}$$

We simulated direct RB under the above error model, for linear and all-to-all connectivity, with both α and ϵ varied.[6] We estimated r_Ω from the data, and compared it to ϵ_Ω. The results of these simulations are summarized in Fig. 6, which quantifies the accuracy of the direct RB error rate r_Ω by the relative error $\delta = (r_\Omega - \epsilon_\Omega)/\epsilon_\Omega$. For $\epsilon_\Omega \leq 1\%$ we see that direct RB is accurate (i.e., $\delta \approx 0$) for both qubit connectivities and even when all the error is on a single qubit ($w_1 = 1$). This is because even when the error is all on one qubit, the random circuit layers almost certainly delocalize any error before another error can occur (if $\epsilon_\Omega = 1\%$ an error is expected to occur approximately every 100 gates, and an error is delocalized in many fewer than 100 gates, as shown in Fig. 5). As the error rate increases up to $\epsilon_\Omega = 10\%$, we observe that r_Ω begins to underestimate ϵ_Ω, and this underestimate is worse with linear connectivity and with greater bias in the error (increasing w_1). This is because, as w_1 increases, sequential errors in a direct RB circuit are more likely to occur on the same qubit, and when ϵ_Ω is large enough an error is not delocalized with high probability before another error occurs on the same qubit. These results therefore validate our theory of direct RB.

4.9 The impact of stabilizer state preparation and measurement errors

The theory in this section has so far assumed perfect SSPAM (although the simulations, above, do not). We now briefly consider the effect of imperfect SSPAM—which is unavoidable in practice. The main impact of errors in the SSPAM is to decrease the $d = 0$ success probability (S_0). Preparing a typical stabilizer state from a computational basis state requires $O(n^2/\log n)$ one- and two-qubit gates [25, 26, 27], so errors in the SSPAM operations are the main limitation on the number of qubits that direct RB (on Clifford gates) can be applied to, as shown by the simulations in Fig. 2. To confirm that SSPAM errors do not have a significant impact on the estimated direct RB error rate, we simulated direct RB under the same error model, but with or without SSPAM error. Specifically, SSPAM was either implemented by compiling them into circuits containing imperfect native gates, or SSPAM was implemented without error.

We simulated direct RB on n qubits with $n = 0, 1, \ldots, 5$ and with 120 randomly-generated stochastic Pauli error models for each n. We used a gate set constructed from parallel applications of $X(\pi/2)$,$Y(\pi/2)$, I, and CNOT gates. For error models with $n \geq 2$, the Pauli error rates in the model are sampled such that the expected two-qubit gate infidelity is q for 120 evenly-spaced values $q \in [0.004, 0.024]$, and each single-qubit gate has an infidelity of $0.1q$. For error models with $n = 1$, the Pauli error rates on each single-qubit gate are sampled so that the expected infidelity of each gate is q. Figure 7 shows the results of these simulations. We observe no

[6] For each connectivity and each value of $\tilde{\epsilon}$ and α we independently simulated a direct RB experiment 50 times, and each direct RB experiment consisted of 30 circuits at benchmark depths $d = 0, 1, 2, 4, 8 \ldots, 128$. In each simulated direct RB experiment we compute the estimate of r_Ω, and then we take the mean of these 50 values.

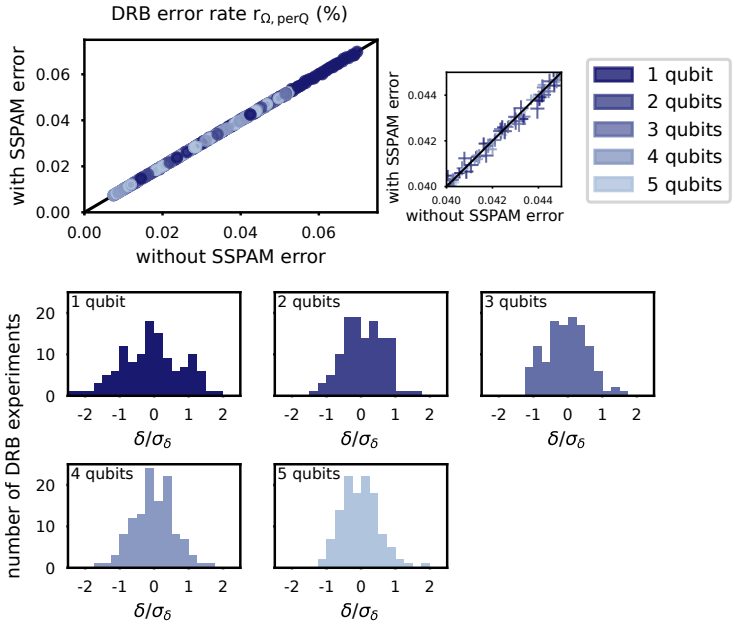

Figure 7: Numerical simulations demonstrate that the direct RB error rate is unaffected by errors in the stabilizer state preparation and measurement (SSPAM) subcircuits, including the error in the intrinsic state preparation and measurement (SPAM). (Top) Each point shows the estimates for the direct RB error rate per qubit $r_{\Omega,\mathrm{perQ}} = 1 - (1 - r_\Omega)^{1/n}$ obtained when direct RB is simulated under the same error model with imperfect SSPAM (horizontal axis) or with artificially perfect SSPAM (vertical axis). Each data point corresponds to a different stochastic Pauli error model with randomly sampled error rates (details in main text). (Bottom) The relative error δ in the direct RB error rate with respect to the direct RB error rate with perfect SSPAM circuits, divided by its standard deviation σ_δ (the standard deviation is calculated using a nonparametric bootstrap). These histograms exhibit no statistically significant evidence for any difference between r_Ω with SSPAM error and r_Ω without SSPAM error.

systematic difference in the direct RB error rate with and without SSPAM error (see upper plots in Fig. 7). Moreover, there is no statistically significant evidence of any difference between r_Ω with or without SSPAM error (see the histograms in Fig. 7). This is evidence that r_Ω is not significantly affected by physically realistic SSPAM errors. This is consistent with the theory for direct RB that we present in Section 6.

4.10 A heuristic theory for direct RB with general errors

In this section we have presented a theory that explains how and why direct RB works when applied to gates that experience stochastic Pauli errors. We now explain how the above theory can be extended to a *descriptive* theory for direct RB with general Markovian errors. The fundamental idea is that layers of random gates decohere errors: coherent errors can systematically add or cancel across short distances in a random circuit, but (on average) they cannot systematically add or cancel across many layers of a random circuit. This can be most easily formalized for circuits containing layers of random single qubit gates, as assumed throughout this section. This is because a single layer of uniformly random single-qubit Clifford gates (an L_1 layer) is sufficient to project *any* error map (a CPTP superoperator) onto the space spanned by tensor-products of local depolarizing channels [85]. So, for direct RB circuits that contain layers of uniformly random single-qubit Clifford gates (or Pauli gates), the behavior of direct RB under any Markovian error model can be modeled using its "equivalent" stochastic Pauli error model—i.e., the model in which

each error map is replaced with its Pauli-twirl. Therefore, the theory in this section also arguably proves that direct RB is reliable for general Markovian errors, although it does not make precise mathematical predictions in that case. This complements the *predictive* theory for direct RB of gates that experience general Markovian errors that we present in Section 6. That theory does not require the above assumption about the distribution of the random layers in the direct RB circuits (i.e., uniformly random L_1 layers), but (unlike the theory in this section) it does not prove that direct RB is reliable when $n \gg 1$ except with very stringent conditions on the gate's error rates.

5 Sequence-asymptotic unitary designs

Our theory for direct RB with general Markovian errors (Section 6) relies on the concept of *sequence-asymptotic unitary 2-designs*. In this section we define sequence-asymptotic unitary 2-designs and we prove some results about their properties. Informally, a sequence-asymptotic unitary 2-design is a gate set $\mathbb{G}$ and a sampling distribution Ω over $\mathbb{G}$ such that the set of unitaries created by length l sequences of gates sampled from Ω converges to a unitary 2-design as $l \to \infty$. The concept of sequence-asymptotic unitary 2-designs is related to approximate unitary 2-designs [45, 44], and the broader theory of convergence to unitary t-designs and other scrambling properties of random circuits [57, 58, 59, 60, 61, 62, 63]. But, because we only use sequence-asymptotic unitary 2-designs as a tool for studying direct RB, we do not discuss the relationship with this prior work.

A plausibly sufficient condition for $(\mathbb{G}, \Omega)$ to constitute a sequence-asymptotic unitary 2-design is that $\mathbb{G}$ generates a unitary 2-design (assuming that Ω has support on all of $\mathbb{G}$). This is because, for a wide class of generating sets and probability distributions, a random sequence of group generators converges to a random group element with increasing length sequences [86, 87, 88]. However, this convergence does not happen for all probability distributions over all generating sets: the long-length distribution can oscillate between the uniform distribution on different parts of the group (an example is given towards the end of this section, and this is only possible if $\mathbb{G}$ is contained within a coset of a proper subgroup of $\mathbb{C}$). In this section, we will show the following: given a set of generators for any group that is a unitary 2-design, length-l sequences of Ω-distributed and independently sampled generators converge to a unitary 2-design as $l \to \infty$, if and only if the set of length-l sequences also generate unitary 2-designs for all l.

5.1 Notation

First we need to define some notation. In this section we consider a quantum system of dimension d, without the restriction to $d = 2^n$ used throughout the rest of this paper. Let $\mathfrak{L}(H_d)$ denote the space of linear operators on the d-dimensional Hilbert space H_d. Let $\mathbb{T}(H_d)$ denote the space of superoperators on H_d, which is the set of all linear operators $\mathcal{S}$ whereby $\mathcal{S} : \mathfrak{L}(H_d) \to \mathfrak{L}(H_d)$. In the following proofs, we will use Hilbert-Schmidt space, and superoperator "stacking". Let

$$\mathrm{B} = \{B_0, B_1, \ldots, B_{d^2-1}\} \tag{48}$$

be an orthonormal basis for $\mathfrak{L}(H_d)$, with respect to the Hilbert-Schmidt inner product

$$\langle X, Y \rangle = \mathrm{Tr}\left(X^\dagger Y\right). \tag{49}$$

We can always choose this basis such that $B_0 = \mathbb{1}/\sqrt{d}$ and $\mathrm{Tr}(B_j) = 0$ for $j > 0$, and this will be assumed herein. For example, with $d = 2^n$ the elements of this basis can be the n-qubit Pauli operators multiplied by $1/\sqrt{2^n}$.

Any $\rho \in \mathfrak{L}(H_d)$ can be expanded with respect to the basis B as

$$\rho = \sum_j \langle B_j, \rho \rangle B_j. \tag{50}$$

As such, ρ may be represented by the column vector

$$|\rho\rangle\!\rangle = \left(\langle B_0, \rho\rangle, \langle B_1, \rho\rangle, \ldots, \langle B_{d^2-1}, \rho\rangle\right)^T. \tag{51}$$

Similarly, a superoperator $\mathcal{X}$ may be represented as the matrix $\mathcal{X}^{\mathrm{HS}}$ with elements $\mathcal{X}^{\mathrm{HS}}_{ij} = \langle B_i, \mathcal{X} B_j \rangle$. In the following we will drop the superscript; we will not use a distinct notation for a superoperator and the matrix representation of the superoperator acting on Hilbert-Schmidt space.

In this and the next section we will be considering linear maps from superoperators to superoperators, which we will refer to as "superchannels" and denote using the script font (e.g., $\mathscr{L}$). When considering superchannels it is often convenient to represent matrices acting on Hilbert-Schmidt space (i.e., superoperators) as vectors in a vector space, and superchannels as matrices acting on that vector space. To achieve this we use the invertible linear "stacking" map

$$\mathrm{vec}\left(|B_j\rangle\!\rangle\langle\!\langle B_k|\right) = |B_k\rangle\!\rangle \otimes |B_j\rangle\!\rangle, \tag{52}$$

which stacks the columns of a matrix. We denote the inverse "unstacking" operation by

$$\mathrm{unvec}\left(|B_k\rangle\!\rangle \otimes |B_j\rangle\!\rangle\right) = |B_j\rangle\!\rangle\langle\!\langle B_k|. \tag{53}$$

From this definition, it follows that

$$\mathrm{vec}(\mathcal{A}\mathcal{B}\mathcal{C}) = \left(\mathcal{C}^T \otimes \mathcal{A}\right) \mathrm{vec}(\mathcal{B}) \tag{54}$$

for Hilbert-Schmidt matrices $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$. Therefore, any superchannel $\mathscr{K}$ with the action $\mathscr{K}(\mathcal{B}) = \sum_i \mathcal{A}_i \mathcal{B} \mathcal{C}_i$ may be rewritten as a matrix

$$\mathrm{mat}[\mathscr{K}] = \sum_i \left(\mathcal{C}_i^T \otimes \mathcal{A}_i\right), \tag{55}$$

that acts on vectors $|\mathcal{B}) \equiv \mathrm{vec}(\mathcal{B})$, where we will use the $|\cdot)$ notation to succinctly denote stacked superoperators while distinguishing them from pure states in Hilbert space (denoted $|\cdot\rangle$) and Hilbert-Schmidt vectors (denoted $|\cdot\rangle\!\rangle$).

5.2 Ordinary unitary designs

Before introducing sequence-asymptotic unitary designs, we first review the definition of an "ordinary" unitary t-design [44, 45]. Although not the original definition [45], an equivalent definition for a unitary t-design is [44]:

Definition 1. *The gate set $\mathbb{G} \subseteq \mathbb{U}$ is a unitary t-design if*

$$\frac{1}{|\mathbb{G}|} \sum_{\mathcal{G} \in \mathbb{G}} \mathcal{G}^{\otimes t} = \int_{\mathcal{U} \in \mathbb{U}} \mathcal{U}^{\otimes t} d\mu \tag{56}$$

where $d\mu$ is the Haar measure on $\mathbb{U}$.

The class of designs important herein are the unitary 2-designs. The definition of a unitary t-design implies that an equivalent characterization of a unitary 2-design is the following: a gate set is a unitary 2-design if and only if, when it "twirls" any channel, it projects this channel onto the space of depolarizing channels. Specifically, define the $\mathbb{G}$-twirl superchannel $\mathscr{T}_{\mathbb{G}} : \mathbb{T}(H_d) \to \mathbb{T}(H_d)$ by

$$\mathscr{T}_{\mathbb{G}}(\mathcal{E}) = \frac{1}{|\mathbb{G}|} \sum_{\mathcal{G}\in\mathbb{G}} \mathcal{G}\mathcal{E}\mathcal{G}^{\dagger}. \tag{57}$$

The twirl superchannel is a linear operator on superoperators. If $\mathbb{G}$ is a unitary 2-design then

$$\mathscr{T}_{\mathbb{G}}(\mathcal{E}) = \mathcal{D}_{\lambda(\mathcal{E})}, \quad \lambda(\mathcal{E}) = \frac{\mathrm{Tr}[\mathcal{E}] - 1}{d^2 - 1}, \tag{58}$$

for any trace-preserving superoperator $\mathcal{E} \in \mathbb{T}(H_d)$ [9], where $\mathcal{D}_{\lambda(\mathcal{E})}$ is a "global" depolarizing channel (i.e, uniform depolarization on the d-dimensional space).

Eq. (58) implies that the twirl superchannel for any unitary 2-design is a rank 2 projector, with support on the depolarizing subspace $\mathbb{S}_{\mathrm{dep}}$ spanned by the identity channel and any non-identity depolarizing channel. Explicitly,

$$\mathbb{S}_{\mathrm{dep}} = \mathrm{span}\left\{\mathcal{D}_0, \mathbb{1}\right\}, \tag{59}$$

where $\mathbb{1}$ denotes the identity superoperator and $\mathcal{D}_0$ is the completely depolarizing channel.

5.3 Sequence-asymptotic unitary designs

We now introduce the notion of a sequence-asymptotic unitary 2-design. This concept is distinct from the "asymptotic unitary 2-designs" of Gross *et al.* [44], which are asymptotically unitary 2-designs with respect to scaling of the dimension d. This definition will be in terms of a set $\mathbb{G}$ of unitaries and a probability distribution Ω over $\mathbb{G}$. Note that we do not assume that $\mathbb{G}$ is finite. Instead, we assume the weaker condition that $\mathbb{G}$ contains a finite number of "gates" that are each either fixed (i.e., a constant matrix) or continuously parameterized. For example, a one-qubit gate set of this sort $\mathbb{G}$ is the Hadamard gate and a continuous family of rotations around σ_z. The probability distribution Ω should be interpreted as an arbitrary weighted mixture of probability distributions over these fixed or parameterized gates. That is, sampling from Ω consists of first selecting one of the gates, and if this selected gate is continuously parameterized we then further sample a value for the continuous parameter(s) according to some distribution. We also assume that all of the probability density functions over any continuous parameters are continuous functions. So when we use the notation $\sum_{\mathcal{G}\in\mathbb{G}}$ this should be interpreted as a sum over the finite number of (perhaps parameterized) gates and an integral over any continuous parameters in each gate.

Definition 2. *Let $\mathbb{G} \subseteq \mathbb{U}$ be a unitary gate set and $\Omega : \mathbb{G} \to [0,1]$ be a probability distribution over $\mathbb{G}$. The tuple $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary t-design if*

$$\lim_{m\to\infty} \left(\sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G}) \mathcal{G}^{\otimes t} \right)^m = \int_{\mathcal{U}\in\mathbb{U}} \mathcal{U}^{\otimes t} d\mu. \tag{60}$$

If $\mathbb{G}$ is a unitary t-design then $(\mathbb{G}, \Pi)$ is also a sequence-asymptotic unitary t-design, where Π is the uniform distribution (this can be shown by induction).

Definition 3. *The $(\mathbb{G}, \Omega)$-twirl superchannel is the linear map $\mathscr{T}_{\mathbb{G},\Omega} : \mathbb{T}(H_d) \to \mathbb{T}(H_d)$ with the action*

$$\mathscr{T}_{\mathbb{G},\Omega}(\mathcal{E}) = \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G}) \mathcal{G}\mathcal{E}\mathcal{G}^{\dagger}. \tag{61}$$

If $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design, Definition 2 implies that the m^{th} power of the $(\mathbb{G}, \Omega)$ twirl superchannel $\mathscr{T}_{\mathbb{G},\Omega}$ converges to a rank 2 projector as $m \to \infty$. As such, the absolute values of the eigenvalues of $\mathscr{T}_{\mathbb{G},\Omega}$ outside of this subspace must be strictly bounded above by 1. The converse is also true: this condition on the eigenvalues of $\mathscr{T}_{\mathbb{G},\Omega}$ is sufficient for any $(\mathbb{G}, \Omega)$ to be a sequence-asymptotic unitary 2-design. We state this more formally in the following proposition, for which we will require the notion of the fixed points of a linear map. The fixed points of a linear map $L : V \to V$, for some vector space V, are those $v \in V$ such that $L(v) = v$. That is, they are the eigenvectors with eigenvalue 1. For example, the space of all fixed points of the twirling map for any unitary 2-design is $\mathbb{S}_{\text{dep}}$ [as defined in Eq. (59)]. The following are formal requirements on properties of the map $\mathscr{T}_{\mathbb{G},\Omega}$, that will allow us to characterize the sequence-asymptotic unitary 2-designs.

Proposition 1. *Let $\mathbb{G} \subseteq \mathbb{U}$ be a unitary gate set and $\Omega : \mathbb{G} \to [0, 1]$ be a probability distribution over $\mathbb{G}$. The tuple $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design if and only if the following two conditions hold:*

1. *The space of all fixed points of $\mathscr{T}_{\mathbb{G},\Omega}$ is the subspace of depolarizing channels $\mathbb{S}_{\text{dep}}$.*
2. *If $\mathscr{T}_{\mathbb{G},\Omega}(\mathcal{E}) = \lambda\mathcal{E}$ with $\mathcal{E} \notin \mathbb{S}_{\text{dep}}$ then $|\lambda| < 1$.*

Proof. We first prove that if $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design then (1) and (2) are implied. If $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design then the fixed points of $\mathscr{T}^m_{\mathbb{G},\Omega}$ converge to $\mathbb{S}_{\text{dep}}$ as $m \to \infty$. The superchannel $\mathscr{T}_{\mathbb{G},\Omega}$ has complex eigenvalues of modulus at most 1. One way to see this is by observing that $\text{mat}\left[\mathscr{T}_{\mathbb{G},\Omega}\right] := \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})\mathcal{G}^* \otimes \mathcal{G}$, which has the same spectrum as $\mathscr{T}_{\mathbb{G},\Omega}$, is a convex sum of unitaries, and so it is a CPTP map. Hence, any fixed point of $\mathscr{T}^m_{\mathbb{G},\Omega}$ is an eigenchannel of $\mathscr{T}_{\mathbb{G},\Omega}$ with an eigenvalue of unit modulus, and so it is also a fixed point of $\mathscr{T}_{\mathbb{G},\Omega}$ if this eigenvalue is 1. Moreover, all of the unit modulus eigenvalues of $\mathscr{T}_{\mathbb{G},\Omega}$ must be equal to 1, because if this was not the case then $\mathscr{T}^m_{\mathbb{G},\Omega}$ would not have a convergent spectrum as $m \to \infty$, which would imply that $(\mathbb{G}, \Omega)$ was not a sequence-asymptotic unitary 2-design. Hence, the space of fixed points of $\mathscr{T}_{\mathbb{G},\Omega}$ is $\mathbb{S}_{\text{dep}}$, and the eigenvalues for all eigenchannels outside of this subspace have absolute value strictly less than 1.

It remains to prove that if (1) and (2) hold then $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design. If (1) and (2) hold this immediately implies that $\mathscr{T}^m_{\mathbb{G},\Omega}$ converges to a rank-2 projector onto $\mathbb{S}_{\text{dep}}$ as $m \to \infty$. Projection onto this subspace is the action of the right hand side of Definition 2 for $t = 2$ [9], so (1) and (2) imply that Definition 2 holds for $t = 2$. □

We now prove a lemma that will be of use later. Here and later, we let $\mathbb{G}^k = \{g_k \dots g_1 \mid g_i \in \mathbb{G}\}$ be the gate set consisting of the set of all length k sequences of gates from $\mathbb{G}$ and Ω^k be the distribution over $\mathbb{G}^k$ obtained by k convolutions of Ω, i.e., $\Omega^k(g_k \dots g_1) = \Omega(g_k) \cdots \Omega(g_1)$.

Lemma 1. *The tuple $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design if and only if, for all $k > 0$, $(\mathbb{G}^k, \Omega^k)$ is a sequence-asymptotic unitary 2-design.*

Proof. Suppose $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design. The composite of twirling maps is a twirling map. Thus by taking the composite map, (1) and (2) of Proposition 1 hold. The reverse direction is trivial. □

Proposition 1 allows us to test whether or not a given $(\mathbb{G}, \Omega)$ tuple is a sequence-asymptotic unitary 2-design, by explicitly calculating the spectrum of $\mathscr{T}_{\mathbb{G},\Omega}$. However, for an n-qubit gate set,

this brute force calculation becomes quickly intractable as n increases—using the stacked Hilbert-Schmidt representation, $\mathscr{T}_{\mathbb{G},\Omega}$ is a $16^n \times 16^n$ dimensional matrix. Moreover, this proposition is not particularly instructive for the task of finding sequence-asymptotic unitary 2-designs. In the following proposition and theorem, we give an alternative characterization of sequence-asymptotic unitary 2-designs.

Proposition 2. *Let $\mathbb{G} \subseteq \mathbb{U}$ be a unitary gate set and $\Omega : \mathbb{G} \to [0,1]$ be a probability distribution over $\mathbb{G}$ with support on all of $\mathbb{G}$. Then $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design if and only if the subspace of $\mathbb{T}(H_d)$ containing all superoperators that are preserved up to a (possibly superoperator-dependent) phase under conjugation by all elements of $\mathbb{G}$ is the space of depolarizing channels $\mathbb{S}_{\text{dep}}$.*

Proof. Assume that $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design. Let $\mathbb{S}$ denote the set of all superoperators preserved under conjugation by all elements of $\mathbb{G}$ up to a phase, meaning the set of all $\mathcal{S} \in \mathbb{T}(H_d)$ satisfying $\mathcal{G}\mathcal{S}\mathcal{G}^{-1} = e^{i\theta_{\mathcal{S}}}\mathcal{S}$ for all $\mathcal{G} \in \mathbb{G}$, where $\theta_{\mathcal{S}}$ is a phase that depends on the superoperator being conjugated. If the superoperator $\mathcal{S}$ is preserved under conjugation up to a phase by all elements of $\mathbb{G}$ then it is a unit-modulus eigenvector of the $(\mathbb{G}, \Omega)$-twirl superchannel $\mathscr{T}_{\mathbb{G},\Omega}$. Thus, by (2) in Proposition 1, it must have $\theta_{\mathcal{S}} = 0$ because $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design by assumption. Hence, $\mathbb{S}$ is a subspace of the set of all fixed points of $\mathscr{T}_{\mathbb{G},\Omega}$. If $\mathbb{G}_\Omega$ is a sequence-asymptotic unitary 2-design then Proposition 1 states that the space of all fixed points of $\mathscr{T}_{\mathbb{G},\Omega}$ is $\mathbb{S}_{\text{dep}}$. Therefore, as all depolarizing maps are preserved under conjugation by any unitary gate, if $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design then $\mathbb{S} = \mathbb{S}_{\text{dep}}$.

Next, we must show that $\mathbb{S} = \mathbb{S}_{\text{dep}}$ implies that $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design. Assume that $\mathbb{S} = \mathbb{S}_{\text{dep}}$, and consider $\mathcal{V} \in \mathbb{S}$ such that $\mathbb{G}\mathcal{V}\mathbb{G}^{-1} = e^{i\theta_{\mathcal{V}}}\mathcal{V}$. By definition, $\mathcal{V} \in \mathcal{S}$ and hence by assumption, $\mathcal{V} \in \mathbb{S}_{\text{dep}}$. If $\mathcal{V}$ has support on the identity superoperator, we find that $\theta_{\mathcal{V}} = 0$ since the identity commutes with all elements of $\mathbb{G}$. The only element in $\mathbb{S}_{\text{dep}}$ without support on the identity superoperator is the completely depolarizing map. However, this as well cannot have a nontrivial phase, as that would imply that channels that are mixtures of the identity superoperator and the completely depolarizing channel would be required to have the same nontrivial phase, which we have just concluded is impossible. Hence, if $\mathcal{V}$ is preserved up to phase under conjugation by all elements of $\mathbb{G}$, it is preserved with a phase of 1.

Now, consider the $(\mathbb{G}, \Omega)$-twirl superchannel $\mathscr{T}_{\mathbb{G},\Omega}$, and consider any $\mathcal{V}$ satisfying $\mathscr{T}_{\mathbb{G},\Omega}(\mathcal{V}) = e^{i\theta}\mathcal{V}$ for some $e^{i\theta}$. For clarity, let us express this as a matrix acting on a vector, by using the "stack" operation. We then have

$$\sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})[\mathcal{G}^* \otimes \mathcal{G}] \operatorname{vec}(\mathcal{V}) = e^{i\theta} \operatorname{vec}(\mathcal{V}) \tag{62}$$

where the matrix acting on $\operatorname{vec}(\mathcal{V})$ is a convex sum of unitaries, or a weighted sum and integral over unitaries with the total "weighting" integrating to 1. Because unitaries preserve the length of vectors, this can only hold if $[\mathcal{G} \otimes \mathcal{G}] \operatorname{vec}(\mathcal{V}) = e^{i\theta} \operatorname{vec}(\mathcal{V})$ for all $\mathcal{G} \in \mathbb{G}$ for which $\Omega(\mathcal{G}) > 0$, and this is all $\mathcal{G} \in \mathbb{G}$ by the conditions of the proposition. Hence, returning to the unstacked representation, we have that

$$\mathcal{G}\mathcal{V}\mathcal{G}^{-1} = e^{i\theta}\mathcal{V} \tag{63}$$

for all $\mathcal{G} \in \mathbb{G}$. But we have shown above that, if $\mathbb{S} = \mathbb{S}_{\text{dep}}$, this holds if and only if $\mathcal{V} \in \mathbb{S}_{\text{dep}}$ and $e^{i\theta} = 1$. As such, any eigen-operator of $\mathscr{T}_{\mathbb{G},\Omega}$ outside of the depolarizing subspace has an eigenvalue with absolute value strictly less than 1 in modulus. Proposition 1 then implies that $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design. □

Proposition 2 implies the following theorem:

Theorem 1. *Let $\mathbb{G} \subseteq \mathbb{U}$ be a unitary gate set that either generates a group or generates a dense subset of a group, $\mathbb{C}$, and let $\Omega : \mathbb{G} \to [0, 1]$ be a probability distribution over $\mathbb{G}$ with support on all of $\mathbb{G}$. Let $\mathbb{G}^k = \{g_1 \dots g_k \mid g_i \in \mathbb{G}\}$, which generates a group, or a dense subset of a group, $\mathbb{C}^k$. Then $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design if and only if, for all k, $\mathbb{C}^k$ is a unitary 2-design.*

Proof. First we show that if each $\mathbb{C}^k$ is a unitary 2-design, then $\mathbb{G}$ is a sequence-asymptotic unitary 2-design. If $\mathbb{G}$ generates a unitary 2-design, then the depolarizing subspace is a subspace of the space preserved up to a phase by $\mathbb{G}$. Suppose there is some element in the preserved space that is not depolarizing. There is a k-string (a length k sequence of $G \in \mathbb{G}$) that brings this phase arbitrarily close to one. Then, in the group generated by these k-strings, this element is also preserved by $\mathbb{C}^k$ with a phase arbitrarily close to 1. However, we also have that $\mathbb{C}^k$ is a unitary 2-design and thus element must also belong to the depolarizing subspace, and thus we have a contradiction.

Next, we prove that if there is some $\mathbb{C}^k$ that is not a unitary 2-design, then $\mathbb{G}$ is not a sequence-asymptotic unitary 2-design. Let $\mathbb{S}_{\mathbb{G}^k}$ and $\mathbb{S}_{\mathbb{C}^k}$ denote the space of all superoperators preserved under conjugation up to a phase by all elements of $\mathbb{G}^k$ and $\mathbb{C}^k$, respectively. Because $\mathbb{G}^k$ generates $\mathbb{C}^k$, or generates a dense subset of $\mathbb{C}^k$, then $\mathbb{S}_{\mathbb{G}^k} = \mathbb{S}_{\mathbb{C}^k}$. If $\mathbb{C}^k$ is not a unitary 2-design then since $\mathbb{S}_{\mathbb{C}^k} = \mathbb{S}_{\mathbb{G}^k}$, by Proposition 2 the elements of $\mathbb{S}_{\mathbb{G}^k}$ preserve some superoperator up to a phase that is not a depolarizing channel and thus $\mathbb{S}_{\mathbb{G}^k} \neq \mathbb{S}_{\text{dep}}$. But if $\mathbb{S}_{\mathbb{G}^k} \neq \mathbb{S}_{\text{dep}}$, then $\mathbb{S}_{\mathbb{C}^k} \neq \mathbb{S}_{\text{dep}}$. Therefore, $(\mathbb{G}^k, \Omega^k)$ is not a sequence-asymptotic unitary 2-design, and hence by Lemma 1, $(\mathbb{G}, \Omega)$ is not a sequence-asymptotic unitary 2-design. □

Theorem 1 characterizes the class of all gate sets, and probability distributions over those gate sets, that form sequence-asymptotic unitary 2-designs. Perhaps surprisingly, a necessary and sufficient condition is that the generated group by $\mathbb{G}^k$ is a unitary 2-design for every k. As an example of a consequence of this theorem, consider the gate set $\mathbb{G} = \{\sqrt{Z}H, Z\sqrt{Z}H\}$ for which both elements have a period of 3. $\mathbb{G}$ generates a subgroup of the Clifford group that is a unitary 2-design. However, $\mathbb{G}^3 = \{I, X, Y, Z\}$ is the Pauli group, which does not form a 2-design. Hence, $\mathbb{G}$ does not induce a sequence-asymptotic unitary 2-design.

In any instance where a gate set $\mathbb{G}$ does generate a unitary 2-design but does not induce a sequence-asymptotic unitary 2-design (for any Ω), $\mathbb{G}$ can be altered so that it can induce a sequence-asymptotic unitary 2-design by adding a single element to $\mathbb{G}$. In particular, any gate set $\mathbb{G}$ that generates a unitary 2-design and contains the identity gate is a sequence-asymptotic unitary 2-design. This is a practically reasonable and easy-to-verify condition on a gate set.

6 A theory of Direct RB with gate-dependent errors

In this section we present two closely-related theories for direct RB under general Markovian errors that build on (and are of similar rigor to) the most accurate theories of Clifford group RB under general gate-dependent errors [21, 22, 64, 65]. These theories can be understood in terms of Fourier transforms on functions over groups (as in Ref. [64]), but we chose not to explicitly use this particular mathematical framework. The theories in this section show that the direct RB decay is an exponential ($S_d \approx A + Bp^d$) for sufficiently small gate errors, and that the direct RB error rate (r_Ω) is equal to a gauge-invariant definition for the Ω-weighted average infidelity (ϵ_Ω) of the benchmarked gate set. The main assumptions we make in this section (see also Table 1) are as follows:

1. All errors are Markovian, i.e., a gate's error can be represented by a CPTP map on n qubits.

2. All errors are small ($\epsilon \ll 1$).

As we will see, the maximum value of ϵ for which this theory holds (and proves reliability of direct RB) is a subtle function of the gate set, the number of qubits, and the sampling distribution. However, for few-qubit gate sets (up to $n \sim 4$) and practically-relevant choices for the gate set and sampling distribution (e.g., the uniform distribution over a gate set consisting of $\pi/2$ rotations around the X and Y axes), the conditions on ϵ are physically reasonable.

This section begins with some preliminaries: in Section 6.1 we introduce some notation and three additional more technical and less important assumptions that are used in this section but not elsewhere in the paper; and in Section 6.2 we review the concept of gauge, and gauge-invariant metrics. In Section 6.3 we introduce the first of the two closely-related theories for direct RB that we present in this section. This theory—which we refer to as the *$\mathscr{R}$-matrix theory* for direct RB—combines the regular representation of the group $\mathbb{C}$ with the superoperators for the imperfect gates to construct a matrix $\mathscr{R}$ that can be used to exactly compute the direct RB decay (S_d).

In Sections 6.4-6.7 we then present our second theory for direct RB with general gate-dependent errors—which we refer to as the *$\mathscr{L}$-matrix theory* for direct RB—that enables us to show that $r_\Omega \approx \epsilon_\Omega$. This theory combines the superoperator representation of the group $\mathbb{C}$ with the superoperators for the imperfect gates to construct a matrix $\mathscr{L}$ that is an imperfect version of the twirling superchannel that appears in our theory of sequence-asymptotic unitary 2-designs (Section 5). In Section 6.4 we show how S_d can be approximated by a function of the d^{th} power of this $\mathscr{L}$ matrix. In Section 6.5 we then show how S_d can be expanded into a linear combination of many exponentials, using the spectral decomposition of $\mathscr{L}$. In Section 6.6 we use the properties of sequence-asymptotic unitary 2-designs (Section 5) to show that all but two terms in this linear combination are negligible, i.e., $S_d \approx A + B\gamma^d$ where γ is an eigenvalue of $\mathscr{L}$ that is close to 1. This implies that direct RB approximately measures $r_\gamma = (4^n - 1)(1 - \gamma)/4^n$, i.e., $r_\Omega \approx r_\gamma$. In Section 6.7 we conclude our theory, by building on a proof of Wallman's [22] to show that r_γ is equal to a gauge-invariant definition for the Ω-weighted average infidelity (ϵ_Ω) of the benchmarked gate set. Finally, in Section 6.8 we use the exact $\mathscr{R}$-matrix theory for direct RB, and numerical simulations, to validate our approximate $\mathscr{L}$-matrix theory for direct RB—by demonstrating that $|(r_\Omega - r_\gamma)/r_\gamma|$ is small for a selection of randomly generated error models on a single-qubit gate set. The theories presented in this section have a close relationship to the Fourier transforms over a group, as is the case for Clifford group RB [64].

6.1 Additional assumptions and notation

In addition to the two key assumptions stated above, we will assume that:

3. The state preparation and measurement components of direct RB circuits contain only gates from the benchmarked gate set ($\mathbb{G}$). This is without loss of generality, as the sampling distribution Ω need not have support on all of $\mathbb{G}$.

4. Unconditional compilation of the initial group element and final inversion group element.

5. No randomization of the output bit string.

These assumptions mean that the state preparation circuit implements a random element of $\mathbb{C}$, and the measurement preparation subcircuit implements the unique element in $\mathbb{C}$ that inverts the preceding circuit (as in Clifford group RB). These assumptions are not necessary for much of theory in this section, and we suspect (although have not proven) that they could be discarded. However, they greatly simplify notation.

Because this section is heavy on notation, we simplify some of the notation used elsewhere in this paper. Unlike elsewhere in this paper, we will use $G \in \mathbb{G}$ and $C \in \mathbb{C}$ to represent group elements [elsewhere we use, e.g., $U(G)$] rather than instructions to implement logic operations. Throughout this paper, we denote the superoperator representing the imperfect [perfect] implementation of $G \in \mathbb{G}$ by $\tilde{\mathcal{G}}(G)$ [$\mathcal{G}(G)$]. Similarly, in this section we will denote the superoperator representing the imperfect [perfect] implementation of $C \in \mathbb{C}$ (which are used in the state preparation and measurement stages of direct RB circuits) by $\tilde{\mathcal{C}}(G)$ [$\mathcal{C}(G)$]. Note that $\mathcal{C}(\cdot)$ is a representation of the group $\mathbb{C}$. We will assume a low-error gate set, so that each $\tilde{\mathcal{G}}(G)$ is a small deviation from $\mathcal{G}(G)$ (i.e., $\tilde{\mathcal{G}}(G) \approx \mathcal{G}(G)$) for all $G \in \mathbb{G}$ and similarly $\tilde{\mathcal{C}}(C) \approx \mathcal{C}(C)$ for all $C \in \mathbb{C}$. We will quantify this assumption when used.

6.2 Background: gauge-invariant metrics

In this section we will show that the direct RB error rate (r_Ω) is related to a gauge-invariant version of the average infidelity of an n-qubit gate sampled from Ω (ϵ_Ω). We will therefore first review the concept of gauge in gate sets, as well as gauge-variant and gauge-invariant properties of a gate set. The sets of all as-implemented and ideal operations in direct RB circuits can be represented by

$$\mathbb{G}_{\text{all}} = \{\langle\!\langle x|\}_{x\in\mathbb{B}_n} \cup \{\mathcal{G}(G)\}_{G\in\mathbb{G}} \cup \{\mathcal{C}(C)\}_{C\in\mathbb{C}} \cup \{|0^n\rangle\!\rangle\}\,, \tag{64}$$

$$\tilde{\mathbb{G}}_{\text{all}} = \{\langle\!\langle E_x|\}_{x\in\mathbb{B}_n} \cup \{\tilde{\mathcal{G}}(G)\}_{G\in\mathbb{G}} \cup \{\tilde{\mathcal{C}}(C)\}_{C\in\mathbb{C}} \cup \{|\rho\rangle\!\rangle\}\,, \tag{65}$$

respectively, where $\mathbb{B}_n$ is the set of all length n bit strings, E_x is a measurement effect representing the x measurement outcome (so $\langle\!\langle E_x| \approx \langle\!\langle x|$), and ρ represents imperfect state initialization (so $|\rho\rangle\!\rangle \approx |0^n\rangle\!\rangle$). (Furthermore, due to our assumption in this section that the target bit string is always the all-zeros bit string, we only need the $x = 0^n$ effect).

All ideal and actual outcomes of circuits in direct RB can be computed from $\mathbb{G}_{\text{all}}$ and $\tilde{\mathbb{G}}_{\text{all}}$, respectively. However, the predictions of these models are unchanged under a gauge transformation—all operation sets of the form

$$\tilde{\mathbb{G}}_{\text{all}}(\mathcal{M}) = \left\{\langle\!\langle E_x|\mathcal{M}^{-1}\right\}_{x\in\mathbb{B}_n} \cup \left\{\mathcal{M}\tilde{\mathcal{G}}_i\mathcal{M}^{-1}\right\} \cup \{\mathcal{M}|\rho\rangle\!\rangle\}\,, \tag{66}$$

make identical predictions for the outcome distributions of all circuits, where $\mathcal{M}$ is any invertible linear map, called a gauge transformation [1, 2]. If f is a property of $\tilde{\mathbb{G}}_{\text{all}}$, it is only experimentally observable if

$$f\left(\tilde{\mathbb{G}}_{\text{all}}\right) = f\left(\tilde{\mathbb{G}}_{\text{all}}(\mathcal{M})\right)\,, \tag{67}$$

for all $\mathcal{M}$, i.e., f must be gauge-invariant. Many metrics of error for a gate or set of gates that are defined as functions of process matrices are not gauge-invariant [1, 2, 21]. This includes the [in]fidelity of a superoperator, so ϵ_Ω is not gauge-invariant. This implies that an RB error rate (which is observable) cannot be equal to the standard, gauge-variant definition of the average infidelity of a gate set [21], i.e., to connect RB error rates to infidelities we need a gauge-invariant definition for infidelity [21, 22, 65].

6.3 An exact theory of direct randomized benchmarking

Here we present a theory for direct RB that is exact, generalizing a theory for Clifford group RB presented in [21]. This theory, which we call the $\mathscr{R}$-matrix theory for direct RB, provides an exact formula for the average success probability (S_d), as a function of an imperfect gate set $\tilde{\mathbb{G}}$ and a sampling distribution Ω. This formula is efficient in d (but not n) to compute. We will find this theory useful for verifying the accuracy of the approximate theory (that shows that $r_\Omega \approx \epsilon_\Omega$)

introduced later in this section. Our exact theory requires one additional assumptions: we require that $\mathbb{G}$ generates a finite group $\mathbb{C}$.

Our theory starts from the formula for S_d that follows directly from the definition of S_d and the direct RB circuits. Specifically,

$$S_d = \langle\!\langle E_0 | \tilde{\mathcal{S}}_d | \rho \rangle\!\rangle, \tag{68}$$

where

$$\tilde{\mathcal{S}}_d = \sum_{G_d \in \mathbb{G}} \cdots \sum_{G_1 \in \mathbb{G}} \sum_{C \in \mathbb{C}} [\Omega(G_d) \cdots \Omega(G_1) \Pi(C) \tilde{\mathcal{C}}(C^{-1} G_1^{-1} \cdots G_d^{-1}) \tilde{\mathcal{G}}(G_d) \cdots \tilde{\mathcal{G}}(G_1) \tilde{\mathcal{C}}(C)], \tag{69}$$

where (as above) Π is the uniform distribution. To expresses $\tilde{\mathcal{S}}_d$ (and so S_d) in an efficient-in-d form, we use the regular representation $[R(\cdot)]$ of the group $\mathbb{C}$. To specify these matrices explicitly, we index the elements of $\mathbb{C}$ by $i = 1, 2, \ldots, |\mathbb{C}|$, with $i = 1$ corresponding to the identity group element. Then $R(G)$ is the $\mathbb{C} \times \mathbb{C}$ matrix with elements given by $R(G)_{jk} = 1$ if $GC_k = C_j$ and $R(G)_{jk} = 0$ otherwise. So, $R(G)$ is a permutation matrix in which the i^{th} row encodes the product of G and the i^{th} group element. We now use the regular representation to define a matrix $[\mathscr{R}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}]$ that can be used to model sequences of gates from $\tilde{\mathbb{G}}$ that are independent samples from Ω. We define $\mathscr{R}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ by

$$\mathscr{R}_{\mathbb{G},\tilde{\mathbb{G}},\Omega} = \sum_{G \in \mathbb{G}} \Omega(G) R(G) \otimes \tilde{\mathcal{G}}(G). \tag{70}$$

We now express $\tilde{\mathcal{S}}_d$ in terms of $\mathscr{R}$ matrices, and to do so we will need the vector

$$\vec{v} = (1, 0, 0, \ldots)^T \otimes I, \tag{71}$$

where I is the $4^n \times 4^n$ dimensional identity superoperator (and the first entry of $\vec{v}$ corresponds to the identity group element). We can express $\tilde{\mathcal{S}}_d$ in terms of $\mathscr{R}$ matrices and $\vec{v}$ as follows:

$$\tilde{\mathcal{S}}_d = \vec{v}^T \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \mathscr{R}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega} \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \vec{v}, \tag{72}$$

noting that $\mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$ represents the application of a uniformly random element from $\tilde{\mathbb{C}}$. To understand why this equation holds, consider

$$\vec{w} = \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \mathscr{R}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega} \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \vec{v}. \tag{73}$$

This is a vector of superoperators where the i^{th} element of $\vec{w}$ consists of an average over all sequences of superoperators consisting of (1) a uniformly random element of $\tilde{\mathbb{C}}$, (2) d elements of $\tilde{\mathbb{G}}$ sampled from Ω, and (3) a uniformly random element of $\tilde{\mathbb{C}}$, *conditioned* on the overall ideal (error-free) action of the sequence being equal to the group element $\mathcal{C}(C_i)$. To average over all direct RB sequences of length d we simply need to condition on the overall action of the sequence being the identity—which is $\vec{v}^T \vec{w}$.

To obtain our final result—a formula for S_d—we substitute Eq. (72) into Eq. (68) to obtain

$$S_d = \langle\!\langle E_0 | \vec{v}^T \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \mathscr{R}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega} \mathscr{R}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \vec{v} \, | \rho \rangle\!\rangle. \tag{74}$$

This formula for S_d is efficient in d to calculate (simply via multiplication of $\mathscr{R}$ matrices, or via the spectral decomposition of the two $\mathscr{R}$ matrices). The $\mathscr{R}$ matrices grow quickly with both the size of the generated group (i.e., with $|\mathbb{C}|$) and n. However, the $\mathscr{R}$ matrices are sufficiently small when $\mathbb{C}$ is the single-qubit Clifford group (they are 96×96 matrices) to use this theory

to numerically evaluate S_d (without resorting to sampling) for any single-qubit gate set $\mathbb{G}$ that generates the Clifford group. As such, this theory is useful for numerically studying direct RB in the single-qubit setting—and this will allow us to validate our theory that shows that $r_\Omega \approx \epsilon_\Omega$ (see Section 6.8).

Equation (74) can be rewritten in terms of the d^{th} power of the eigenvalues of $\mathscr{R}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. However, this does not immediately imply that $S_d \approx A + Bp^d$ (each $\mathscr{R}$ matrix has $4^n \times |\mathbb{C}|$ eigenvalues).

6.4 Modelling the direct RB decay with twirling superchannels

In this and the subsequent subsections, we present our *$\mathscr{L}$-matrix theory* for direct RB. In this theory, we will show how the direct RB decay, and error rate, can be approximately expressed in terms of the following matrix

$$\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega} = \sum_{G\in\mathbb{G}} \Omega(G)\mathcal{G}(G) \otimes \tilde{\mathcal{G}}(G). \tag{75}$$

Like the $\mathscr{R}$ matrix [Eq. (70)], this $\mathscr{L}$ matrix consists of summing over a tensor product of (1) a representation of the group $\mathbb{C}$ generated by $\mathbb{C}$ and (2) the as-implemented superoperators. In the case of the $\mathscr{R}$ matrix, this representation of the group was the regular representation $[R(\cdot)]$, and in the case of the $\mathscr{L}$ matrix it is the standard superoperator representation $[\mathcal{G}(\cdot)]$ of the group.

To derive our theory, we first show how the direct RB success probability (S_d) can be approximately expressed as a function of twirling superchannels $\mathscr{Q}$ whereby $\text{mat}[\mathscr{Q}] = \mathscr{L}$. In subsequent subsections, this will enable us to show that the direct RB decay is approximately exponential, and that $r_\Omega \approx \epsilon_\Omega$. Our theory will use *error maps* for the as-implemented group elements $\tilde{\mathbb{C}}$. Let

$$\Lambda(C) = \mathcal{C}(C)^\dagger \tilde{\mathcal{C}}(C), \tag{76}$$

for $C \in \mathbb{C}$, which is the "pre-gate" error map for the superoperator $\tilde{\mathcal{C}}(C)$. We also use the average error map ($\bar{\Lambda}$), defined by

$$\bar{\Lambda} = \sum_{C\in\mathbb{C}} \Pi(C)\Lambda(C). \tag{77}$$

The first step to deriving our $\mathscr{L}$-matrix theory for direct RB is to approximate the imperfect initial group element using a gate-independent error channel. To do so, we first note that Eq. (69) can be rewritten as

$$\tilde{\mathcal{S}}_d = \sum_{C\in\mathbb{C}} \sum_{G_d\in\mathbb{G}} \cdots \sum_{G_1\in\mathbb{G}} [\Pi(C)\Omega(G_d)\cdots\Omega(G_1)\tilde{\mathcal{C}}(C)\tilde{\mathcal{G}}(G_d)\cdots\tilde{\mathcal{G}}(G_1)\tilde{\mathcal{C}}(\{CG_d\cdots G_1\}^{-1})], \tag{78}$$

i.e., we have rewritten the first group element in terms of the G_i and the inversion group element (which is now denoted by C). Now we have that

$$\tilde{\mathcal{C}}(\{CG_d\cdots G_1\}^{-1}) = \mathcal{G}^\dagger(G_1)\cdots\mathcal{G}^\dagger(G_d)\mathcal{C}^\dagger(C)\Lambda(\{CG_d\cdots G_1\}^{-1}), \tag{79}$$

and by substituting this into Eq. (78), we obtain

$$\tilde{\mathcal{S}}_d = \sum_{C\in\mathbb{C}} \sum_{G_d\in\mathbb{G}} \cdots \sum_{G_1\in\mathbb{G}} [\Pi(C)\Omega(G_d)\cdots\Omega(G_1)\tilde{\mathcal{C}}(C)\tilde{\mathcal{G}}(G_d)\cdots\tilde{\mathcal{G}}(G_1) \times \mathcal{G}^\dagger(G_1)\cdots\mathcal{G}^\dagger(G_d)\mathcal{C}^\dagger(C)\Lambda((\{G_d\cdots G_1\}^{-1})]. \tag{80}$$

By replacing the error maps in this equation (i.e., the error maps for each initial group element) with their average, we obtain

$$\tilde{\mathcal{S}}_d = \sum_{C\in\mathbb{C}} \sum_{G_d\in\mathbb{G}} \cdots \sum_{G_1\in\mathbb{G}} [\Pi(C)\Omega(G_d)\cdots\Omega(G_1)\tilde{\mathcal{C}}(C)\tilde{\mathcal{G}}(G_d)\cdots\tilde{\mathcal{G}}(G_1) \quad \times \\ \mathcal{G}^\dagger(G_1)\cdots\mathcal{G}^\dagger(G_d)\mathcal{C}^\dagger(C)]\bar{\Lambda} + \Delta_d, \quad (81)$$

where Δ_d absorbs the approximation error. This approximation breaks the dependencies, enabling independent averaging, as we can rearrange Eq. (81) to

$$\tilde{\mathcal{S}}_d = \Bigg\{ \sum_{C\in\mathbb{C}} \Pi(C)\tilde{\mathcal{C}}(C) \Bigg\{ \sum_{G_d\in\mathbb{G}} \Omega(G_d)\tilde{\mathcal{G}}(G_d)\cdots \\ \Big\{ \sum_{G_1\in\mathbb{G}} \Omega(G_1)\tilde{\mathcal{G}}(G_1)\mathcal{G}^\dagger(G_1) \Big\} \cdots \mathcal{G}^\dagger(G_d) \Bigg\} \mathcal{C}^\dagger(C) \Bigg\} \bar{\Lambda} + \Delta_d. \quad (82)$$

This can be more concisely expressed using two superchannels:

$$\tilde{\mathcal{S}}_d = \left(\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \circ \mathscr{Q}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega}\right)[I]\bar{\Lambda} + \Delta_d, \quad (83)$$

where I is the identity superoperator, and $\mathscr{Q}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ and $\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$ are imperfect twirling superchannels, defined by

$$\mathscr{Q}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}[\mathcal{E}] = \sum_{G\in\mathbb{G}} \Omega(G)\tilde{\mathcal{G}}(G)\mathcal{E}\mathcal{G}^\dagger(G), \quad (84)$$

$$\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}[\mathcal{E}] = \sum_{C\in\mathbb{C}} \Pi(C)\tilde{\mathcal{C}}(C)\mathcal{E}\mathcal{C}^\dagger(C). \quad (85)$$

By substituting Eq. (83) into Eq. (68) we obtain

$$S_d = \langle\!\langle E_0 | \left(\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \circ \mathscr{Q}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega}\right)[I] | \rho' \rangle\!\rangle + \delta_d, \quad (86)$$

where $|\rho'\rangle\!\rangle = \bar{\Lambda}|\rho\rangle\!\rangle$ and

$$\delta_d = \langle\!\langle E_0 | \Delta_d | \rho \rangle\!\rangle. \quad (87)$$

Equation (86) represents the average direct RB success probability in terms of imperfect twirling superchannels, which builds on similar results for Clifford group RB [21, 22, 89].

Later we will apply the theory of sequence-asymptotic unitary 2-designs (Section 5) to understand S_d, and to do so we will find it useful to replace the $\mathscr{Q}$ superchannels in Eq. (86) with the closely related $\mathscr{L}$ matrices [see Eq. (75)]. Consider

$$\mathrm{vec}(\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} \circ \mathscr{Q}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega}[I]) = \mathrm{mat}(\mathscr{Q}_{\mathbb{C},\tilde{\mathbb{C}},\Pi})\mathrm{mat}(\mathscr{Q}_{\mathbb{G},\tilde{\mathbb{G}},\Omega})^d | I), \quad (88)$$

where, as introduced in Section 5, $\mathrm{vec}[\cdot]$ denotes the "stacking" map, $\mathrm{mat}[\cdot]$ turns a superchannel into a matrix and $|\cdot)$ turns a superoperator into a vector. So, e.g., we find that

$$\mathrm{mat}[\mathscr{Q}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}] = \sum_{G\in\mathbb{G}} \Omega(G)\mathcal{G}(G) \otimes \tilde{\mathcal{G}}(G) = \mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}, \quad (89)$$

where $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ is the matrix introduced in Eq. (75) [here we have assumed a Hermitian basis for superoperators, so that $\mathcal{G}(G)$'s element are real numbers]. We can therefore rewrite Eq. (86) in terms of two $\mathscr{L}$ matrices:

$$S_d = \langle\!\langle E_0 | \mathrm{unvec}\left[\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}\mathscr{L}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega} | I)\right] | \rho' \rangle\!\rangle + \delta_d. \quad (90)$$

Equation (90) contains an approximation error term (δ_d) and we now bound this term. Using basic properties of CPTP maps and the diamond norm it may be shown that

$$|\delta_d| \leq \|\Delta_d\|_\diamond \leq \sum_{C\in\mathbb{C}} \Pi(C) \left\|\bar{\Lambda} - \Lambda(C)\right\|_\diamond \equiv \Delta, \tag{91}$$

where $\|\cdot\|_\diamond$ is the diamond norm.[7] Error maps are not gauge-invariant, so the size of δ_d (and the size of our upper-bound on $|\delta_d|$) depends on the representation of $\tilde{\mathbb{G}}_{\rm all}$. Even for low-error gates—here meaning that every $\Lambda(C)$ is close to the identity in some representation—Δ and δ_d can always be made large by choosing a "bad" gauge in which to express $\tilde{\mathbb{G}}_{\rm all}$. But our bound holds for any CPTP representation, and so $|\delta_d| \leq \Delta_{\rm min}$ with $\Delta_{\rm min}$ the minimum of Δ over all CPTP representations of $\tilde{\mathbb{G}}_{\rm all}$. In the following, we will assume we are using a representation in which δ_d is small, which always exists with sufficiently low-error gates. Finally, we note that it is possible to improve our bound on δ_d, using an analysis similar to Wallman's [22] treatment of Clifford group RB, which bounds a quantity that is closely related to δ_d from above by a function that decays exponentially in d.

6.5 Modelling the direct RB decay using the spectrum of twirling superchannels

We have shown that the direct RB average success probability (S_d) can be approximated by a function of the $d^{\rm th}$ power of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ [see Eq. (90)]. This implies that S_d can be approximated using the spectral decomposition of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. We now derive the functional form of this spectral decomposition, which in the subsequent subsection we will use to show that $S_d \approx A + Bp^d$. Let $\tilde{\gamma}_i$ for $i = 1, 2, \ldots, 16^n$ denote the eigenvalues of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ ordered by descending absolute value, i.e., $|\tilde{\gamma}_i| \geq |\tilde{\gamma}_{i+1}|$ for all i. Then Eq. (90) implies that S_d may be written as

$$S_d = \tilde{\omega}_1\tilde{\gamma}_1^d + \tilde{\omega}_2\tilde{\gamma}_2^d + \cdots + \tilde{\omega}_{16^n}\tilde{\gamma}_{16^n}^d + \delta_d, \tag{92}$$

where the $\tilde{\omega}_k$ are d-independent parameters. Note that the eigenvalues of an $\mathscr{L}$ matrix are gauge-invariant—because gauge transformations act on a $\mathscr{L}$ matrix by matrix conjugation, with a matrix of the form $I \otimes \mathcal{M}$—so the $\tilde{\gamma}_i$ are gauge-invariant.

We now derive a formula for $\tilde{\omega}_k$. It is convenient to rewrite Eq. (90) as

$$S_d = \langle\!\langle E_0|\tilde{S}'_d|\rho'\rangle\!\rangle + \delta_d, \tag{93}$$

where

$$\tilde{S}'_d = \text{unvec}\left[\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}\mathscr{L}^d_{\mathbb{G},\tilde{\mathbb{G}},\Omega}|I)\right]. \tag{94}$$

We now write $\tilde{S}'_d$ in terms of the eigenvectors and eigenvalues of the two $\mathscr{L}$ matrices, $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ and $\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$. Let $\tilde{\eta}_i$ for $i = 1, 2, \ldots, 16^n$ denote the eigenvalues of $\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$ ordered by descending absolute value, i.e., $|\tilde{\eta}_i| \geq |\tilde{\eta}_{i+1}|$ for all i (noting that $\tilde{\gamma}_i$ denote the ordered eigenvalues of $\mathscr{L}_{\mathbb{G}\tilde{\mathbb{G}},\Omega}$). Moreover, let $\mathcal{L}_{\mathbb{C},\tilde{\mathbb{C}},i}$ and $\mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},i}$ ($\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},i}$ and $\mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},i}$) be superoperators that, when stacked, are mutually orthonormal left and right eigenvectors of the matrix $\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$ (the matrix $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$) with

[7] An explicit derivation that immediately implies this relation is given in [21] in the context of an equivalent relation in Clifford group RB (see the supplemental material in [21], and also see [8, 89] for similar work). Note that this upper-bound is d independent, unlike the similar bound for approximating Clifford group RB decay curves in [8] that grow with d and which is therefore not able to provide useful guarantees on S_d [21, 22].

eigenvalue $\tilde{\eta}_i$ (eigenvalue $\tilde{\gamma}_i$). That is

$$\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}\Big|\mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},i}\Big) = \tilde{\gamma}_i\Big|\mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},i}\Big), \tag{95}$$

$$\Big(\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},i}\Big|\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega} = \Big(\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},i}\Big|\tilde{\gamma}_i, \tag{96}$$

$$\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}\Big|\mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},i}\Big) = \tilde{\eta}_i\Big|\mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},i}\Big), \tag{97}$$

$$\Big(\mathcal{L}_{\mathbb{C},\tilde{\mathbb{C}},i}\Big|\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi} = \Big(\mathcal{L}_{\mathbb{C},\tilde{\mathbb{C}},i}\Big|\tilde{\eta}_i, \tag{98}$$

with

$$\Big(\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},i} \mid \mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},j}\Big) = \delta_{ij}, \quad \Big(\mathcal{L}_{\mathbb{C},\tilde{\mathbb{C}},i} \mid \mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},j}\Big) = \delta_{ij}. \tag{99}$$

Writing the two $\mathscr{L}$ matrices in terms of their eigenvalues and their left and right eigenvectors, and substituting this decomposition of the $\mathscr{L}$ matrices into Eq. (94), we obtain

$$\tilde{\mathcal{S}}'_d = \sum_{jk} \tilde{\eta}_j \tilde{\gamma}_k^d \text{unvec}\left[|\mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},j})(\mathcal{L}_{\tilde{\mathbb{C}},\mathbb{C},j}|\mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},k})(\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},k}||I)\right]. \tag{100}$$

By substituting Eq. (100) into Eq. (93) and comparing the resultant equation to Eq. (92), it follows that

$$\tilde{\omega}_k = \langle\!\langle E'_0|\mathcal{S}'_{d,k}|\rho\rangle\!\rangle \tag{101}$$

where

$$\mathcal{S}'_{d,k} = \sum_j \tilde{\eta}_j \text{ unvec } \left[|\mathcal{R}_{\mathbb{C},\tilde{\mathbb{C}},j})(\mathcal{L}_{\mathbb{C},\tilde{\mathbb{C}},j}|\mathcal{R}_{\mathbb{G},\tilde{\mathbb{G}},k})(\mathcal{L}_{\mathbb{G},\tilde{\mathbb{G}},k}||I)\right]. \tag{102}$$

This formula shows that we can quantify $\tilde{\omega}_k$ by studying the overlap in the eigenchannels of the two $\mathscr{L}$ matrices, as well as the spectrum of $\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$.

6.6 The direct RB decay is approximately exponential

We have now shown that the average success probability (S_d) in direct RB can be written as a linear combination of the d^{th} powers of the 16^n eigenvalues of the $\mathscr{L}$ matrix ($\tilde{\gamma}_k$), and we have derived formula for the coefficients in this sum ($\tilde{\omega}_k$). We now use the theory of sequence-asymptotic unitary 2-designs (Section 5), as well as the theory of ordinary unitary 2-designs, to show that $S_d \approx A + B\gamma^d$ where $\gamma \equiv \tilde{\gamma}_2$ is the second largest eigenvalue of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. Our theory relies on two properties of $\mathbb{G}$, Ω, and $\tilde{\mathbb{G}}$: (1) $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design, and (2) the gates are low error, i.e., $\tilde{\mathbb{G}}$ is close to $\mathbb{G}$ and $\tilde{\mathbb{C}}$ is close to $\mathbb{C}$.

This second condition, above, implies that both the $\mathscr{L}$ matrices that appear in Eq. (94) [$\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ and $\mathscr{L}_{\mathbb{C},\tilde{\mathbb{C}},\Pi}$] will be small perturbations on the $\mathscr{L}$ matrices for perfect, error-free gate sets [i.e., $\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega}$ and $\mathscr{L}_{\mathbb{C},\mathbb{C},\Pi}$]. The group generated by $\mathbb{G}$, i.e, $\mathbb{C}$, is a unitary 2-design. Therefore

$$\mathscr{L}_{\mathbb{C},\mathbb{C},\Pi} = \sum_{C\in\mathbb{C}} \Pi(C)\mathcal{C}(C) \otimes \mathcal{C}(C), \tag{103}$$

is a projection onto the space spanned by a completely depolarizing channel [$\mathcal{D}_0$] and the identity superoperator [I] (see Section 5). Letting η_i denote the ordered eigenvalues of $\mathscr{L}_{\mathbb{C},\mathbb{C},\Pi}$, we have that (1) $\eta_1 = \eta_2 = 1$, with the two-dimensional eigen-space for this eigenvalue spanned by $\mathcal{D}_0$ and I, and (2) $\eta_i = 0$ for $i \geq 3$. Similarly, Propopsition 1 implies that, because $(\mathbb{G}, \Omega)$ is a sequence-asymptotic unitary 2-design,

$$\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega} = \sum_{G\in\mathbb{G}} \Omega(G)\mathcal{G}(G) \otimes \mathcal{G}(G), \tag{104}$$

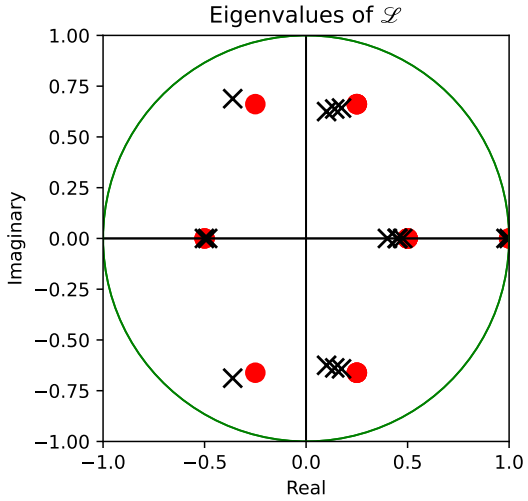

Figure 8: The eigenvalues of the $\mathscr{L}_{\tilde{\mathbb{G}},\mathbb{G},\Omega}$ matrix for the single qubit gate set $\mathbb{G} = \{X(\pi/2), Y(\pi/2)\}$ without errors [red circles], i.e., $\tilde{\mathbb{G}} = \mathbb{G}$, and with a small coherent over-rotation error on each gate [black crosses], i.e. $\tilde{\mathbb{G}} = \{X(\pi/2+\epsilon), Y(\pi/2+\epsilon)\}$, with $\epsilon = 0.1$. There are two unit eigenvalues for the error-free gate set, and all other eigenvalues have absolutely value strictly less than 1. Here the upper spectral gap—meaning the difference in the magnitudes of the second and third largest eigenvalues of $\mathscr{L}$—is $1/2$ for the error-free gate set, and it is still close to $1/2$ for the imperfect gates.

has two unit eigenvalues, which also correspond to the two-dimensional depolarizing subspace, and all other eigenvalues have absolute value strictly less than one. So, letting γ_i denote the ordered eigenvalues of $\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega}$, we have that (1) γ_i $\gamma_1 = \gamma_2 = 1$, with the two-dimensional eigen-space for this eigenvalue spanned by $\mathcal{D}_0$ and I, and (2) $|\gamma_i| \leq 1$ for all $i \geq 3$. Using Eqs. (101)-(102), this then implies that if $\tilde{\mathbb{G}} = \mathbb{G}$ and $\tilde{\mathbb{C}} = \mathbb{C}$, i.e., the gates are perfect, then $\tilde{\omega}_k = 0$ for $k \geq 3$.

Consider now the case of imperfect but low-error gates, i.e., $\tilde{\mathbb{G}} \approx \mathbb{G}$ and $\tilde{\mathbb{C}} \approx \mathbb{C}$. As long as the errors are small, the eigenvalues of $\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega}$ will be close to those of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. This is demonstrated with an example in Fig. 8. Furthermore, as long as the errors are small enough so that the gap between the 2nd largest eigenvalue and the smaller eigenvalues of $\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega}$ does not close, the span of the eigenvectors of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ is a small perturbation on the depolarizing subspace. Similarly, as long as the errors are small enough so that the gap between the 2nd largest eigenvalue and the zero eigenvalues of $\mathscr{L}_{\mathbb{C},\mathbb{C},\Pi}$ does not close, the span of the eigenvectors of $\tilde{\eta}_1$ and $\tilde{\eta}_2$ is a small perturbation on the depolarizing subspace. This then implies that

$$|\tilde{\omega}_k| \ll 1 \quad \forall k \geq 3. \tag{105}$$

Substituting this into Eq. (92), and by noting that 1 is always an eigenvalue of an $\mathscr{L}$ matrix for trace-preserving superoperators (we are assuming CPTP superoperators throughout), we find that

$$S_d = A + B\gamma^d + \delta_d + \delta'_d \tag{106}$$

where $A = \omega_1$, $B = \omega_2$, $\gamma = \tilde{\gamma}_2$ and $\delta'_d = \sum_{k\geq 3}^{16^n} \tilde{\gamma}_k^d \omega_k$ consists of a sum of $16^n - 2$ small terms. Therefore, for small n, we have that

$$S_d \approx A + B\gamma^d, \tag{107}$$

i.e., the direct RB average success probability decay is approximately an exponential (plus a constant).

Interestingly, note that the above theory does not show that $|\delta'_d|$ is small for $n \gg 1$. This is because we have only shown that δ'_d is the sum of $16^n - 1$ small terms. Furthermore, the above theory has limited applicability in the $n \gg 1$ regime for another reason: for any $\mathbb{G}$ consisting of gates constructed from parallel one- and two-qubit gates, the upper spectral gap in the $\mathscr{L}_{\mathbb{G},\mathbb{G},\Omega}$ matrix must decrease in size as n increases. This is because the spectral gap is closely related to convergence to a unitary 2-design, and the number of layers of one- and two-qubit gates needed to implement a good approximation to unitary 2-design increases with n. Yet our less formal theory for direct RB—presented in Section 4—shows that direct RB *is* reliable in the $n \gg 1$ setting, i.e., we do have $S_d \approx A + Bp^d$ for some A, B and p, even when $n \gg 1$. It is an interesting open question as to how to obtain this result from the $\mathscr{R}$ or $\mathscr{L}$ matrix theories presented in this section.

6.7 The direct RB measures a version of infidelity

We have now shown that the direct RB success probability (S_d) is approximately given by $S_d \approx A + B\gamma^d$ where γ is the second largest eigenvalue of the superchannel $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. The direct RB error rate (r_Ω) is defined to be $r_\Omega = (4^n - 1)(1 - p)/4^n$, where p is obtained by fitting S_d to $S_d = A + Bp^d$. The theory presented so far therefore implies that $r_\Omega \approx r_\gamma$ where

$$r_\gamma = (4^n - 1)(1 - \gamma)/4^n. \tag{108}$$

The final part of our theory is to show that r_γ is (exactly) equal to a gauge-invariant version of the mean infidelity of a gate sampled from Ω (ϵ_Ω). That is, we now show that direct RB measures the Ω-weighted average infidelity of the gate set $\tilde{\mathbb{G}}$ it is used to benchmark, in the same sense that Clifford group RB measures the mean infidelity of an implementation of the Clifford group [22, 21].

First we introduce a term for a gate set $\tilde{\mathbb{G}}$ that has sufficiently small errors for the upper spectral gap in the $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ to not close (which is a condition that we required, above, to show that $S_d \approx A + B\gamma^d$).

Definition 4. *Let $(\mathbb{G}, \Omega)$ be a sequence-asymptotic unitary 2-design, and let $\tilde{\mathbb{G}}$ represent some implementation of $\mathbb{G}$. This implementation is $(\mathbb{G}, \Omega)$-benchmarking compatible if the upper spectral gap of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ does not close.*

Next, we present a rather technical proposition, that generalizes a result derived by Wallman [22] (see Theorem 2 therein) for Clifford group RB:

Proposition 3. *Let $(\mathbb{G}, \Omega)$ be a sequence-asymptotic unitary 2-design, and let $\tilde{\mathbb{G}}$ represent an implementation of $\mathbb{G}$ that is sufficiently low error to be $(\mathbb{G}, \Omega)$-benchmarking compatible. Let γ be the second largest eigenvalue of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ by absolute value, and let $\mathcal{E}_1$ and $\mathcal{E}_\gamma$ be eigenoperators of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ with eigenvalues 1 and γ, respectively. Then $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_\gamma$ satisfies*

$$\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E}) = \mathcal{D}_\gamma \mathcal{E}, \tag{109}$$

where $\mathcal{D}_\gamma$ is the depolarization channel with decay constant γ.

Proof. See Appendix A. □

We now use Proposition 3 to prove that $r_\gamma = \epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G})$. Here

$$\epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G}) = \sum_{G \in \mathbb{G}} \Omega(G) \epsilon(\mathcal{E}\tilde{\mathcal{G}}(G)\mathcal{E}^{-1}, \mathcal{G}(G)), \tag{110}$$

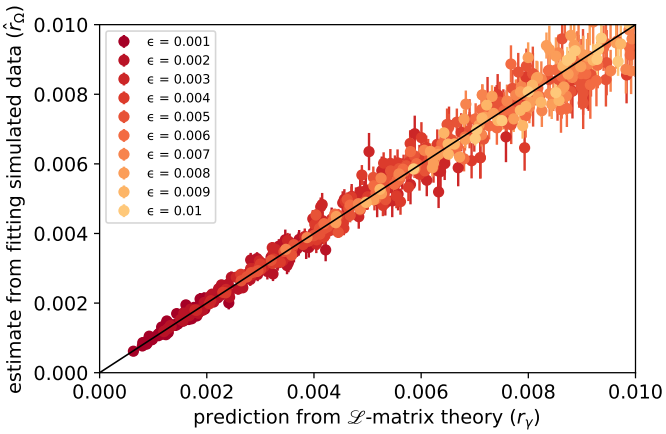

Figure 9: Verifying the accuracy of our $\mathscr{L}$ matrix theory for direct RB, using numerical simulations. We show the theory's prediction for the direct RB error rate (r_γ) versus an estimate $\hat{r}_\Omega$ of the true direct RB error rate (r_Ω) obtained by simulating direct RB. Each data point shows r_γ versus $\hat{r}_\Omega$ for a different randomly generated Markovian error models (see main text for error model details). We find that r_γ and $\hat{r}_\Omega$ are consistent, but note that small systematic deviations between r_γ and r_Ω would not be visible due to the finite sample error on $\hat{r}_\Omega$.

where $\mathcal{E}$ is as defined in Proposition 3. Equation (110) is the Ω-weighted average infidelity between the superoperators in $\mathbb{G}$ and $\tilde{\mathbb{G}}(\mathcal{E})$, where $\tilde{\mathbb{G}}(\mathcal{E})$ expresses the imperfect operations in a particular gauge:[8]

$$\tilde{\mathbb{G}}(\mathcal{E}) := \left\{ \mathcal{E}\tilde{\mathcal{G}}(G)\mathcal{E}^{-1} \mid G \in \mathbb{G} \right\}. \tag{111}$$

Note that the superoperators in $\tilde{\mathbb{G}}(\mathcal{E})$ are not necessarily completely positive (this follows because it is true in the case of Clifford group RB [21, 22], and Clifford group RB is a special case of direct RB). Furthermore, note that $\tilde{\mathbb{G}}(\mathcal{E})$ is ill-defined if $\mathcal{E}$ is singular, and in any such case we take $\tilde{\mathbb{G}}(\mathcal{E})$ to instead be defined using a small perturbation on $\mathcal{E}$. In the following proposition, which is a generalization of a result for Clifford group RB shown in [21, 22], we prove that $r_\gamma = \epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G})$.

Proposition 4. *Let $\mathbb{G}, \Omega, \tilde{\mathbb{G}}, \gamma$ and $\mathcal{E}$ be as defined in Proposition 3. Then $r_\gamma = (d^2-1)(1-\gamma)/d^2$ satisfies*

$$r_\gamma = \epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G}). \tag{112}$$

Proof. From Proposition 3 we have that $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E}) = \mathcal{D}_\gamma\mathcal{E}$, and so $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E})\mathcal{E}^{-1} = \mathcal{D}_\gamma$. Taking the entanglement infidelity to the identity superoperator of both sides of this equality, and using the definition of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$, we have that

$$\epsilon\left(\sum_{G\in\mathbb{G}} \Omega(G)\mathcal{G}^{-1}(G)\mathcal{E}\tilde{\mathcal{G}}(G)\mathcal{E}^{-1}, \mathbb{1} \right) = \epsilon(\mathcal{D}_\gamma, \mathbb{1}) \tag{113}$$

Now using the well-known relation [8]

$$\epsilon(\mathcal{D}_\lambda, \mathbb{1}) = \frac{(4^n-1)(1-\lambda)}{4^n} \tag{114}$$

[8] Note that the $\mathcal{E}$ transformation is also applied to the SPAM whenever any probabilities are to be calculated, but for simplicity we have dropped the SPAM operations from the notation throughout this subsection.

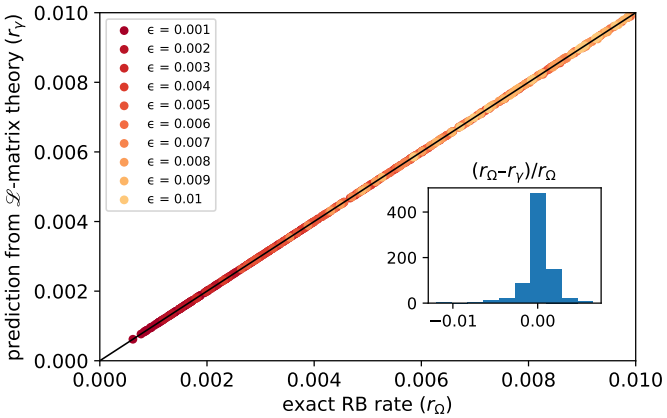

Figure 10: Verifying the accuracy of our $\mathscr{L}$ matrix theory for direct RB, using our exact $\mathscr{R}$ matrix theory. We show the $\mathscr{L}$ matrix theory's prediction for the direct RB error rate (r_γ) versus r_Ω computed from our exact $\mathscr{R}$ matrix theory (the $\mathscr{R}$ matrix theory enables computing S_d exactly, and we then fit S_d to $S_d = A + Bp^d$). Each data point shows r_γ versus r_Ω for the same randomly generated Markovian error models as in Figure 9. The inset shows a histogram of the relative error $[(r_\Omega - r_\gamma)/r_\Omega]$, demonstrating that our $\mathscr{L}$ matrix theory is extremely accurate. We observe extremely close agreement between r_Ω and r_γ, with no more than $\pm 1\%$ relative error (see inset histogram).

and the definition of r_γ, it follows that $\epsilon(\mathcal{D}_\gamma, \mathbb{1}) = r_\gamma$. Using this equality, the linearity of ϵ, and the easily verified relation $\epsilon(B^{-1}A, \mathbb{1}) = \epsilon(A, B)$ for any superoperators A and B with B invertible, Eq. (113) implies that

$$\sum_{G \in \mathbb{G}} \Omega(\mathcal{G}) \epsilon \left(\mathcal{E} \tilde{\mathcal{G}}(G) \mathcal{E}^{-1}, \mathcal{G}(G) \right) = r_\gamma \tag{115}$$

By noting that $\tilde{\mathbb{G}}(\mathcal{E}) = \{\mathcal{E}\tilde{\mathcal{G}}(G)\mathcal{E}^{-1} \mid G \in \mathbb{G}\}$, it follows immediately from the definition of ϵ_Ω, which is the average infidelity of a gate sampled from Ω, that the left hand side of this equation is the $\epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G})$. As such $r_\gamma = \epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G})$, as required. □

We now summarize the results of our theory for direct RB in the following lemma:

Lemma 2. *The direct RB error rate (r_Ω) is approximately equal to the Ω-weighted average infidelity of the benchmarked gate set: $r_\Omega \approx \epsilon_\Omega(\tilde{\mathbb{G}}(\mathcal{E}), \mathbb{G})$.*

6.8 Validating the theory using numerical simulation

In this section we verify the correctness of our $\mathscr{L}$ matrix theory for direct RB, using numerical simulations. We simulated direct RB on a single qubit with a gate set consisting of $X(\pi/2)$ and $Y(\pi/2)$ gates. We choose an error model consisting of random Markovian errors with rates that contribute no more than ε to the gate's infidelity, for some small $\varepsilon > 0$, using the error generator formalism in [90]. That is, the superoperator $\tilde{\mathcal{G}}(G)$ for each gate is modelled as

$$\tilde{\mathcal{G}}(G) = \exp\left(\sum_{i=1}^{12} v_i(G) \mathcal{V}_i \right) \mathcal{G}(G) \tag{116}$$

where the $\mathcal{V}_i$ form a basis for the space of Markovian errors. For this example, the 12 error generators consist of three each of C, A, S and H type generators defined in [90]—where S and H are Pauli stochastic and Hamiltonian error generators respectively, while the remaining "Pauli-correlation" and "active" error generators augment the stochastic errors. For Hamiltonian errors,

the leading order contribution to a gate's infidelity is at second order, whereas for stochastic errors it is at first order. Therefore, the Hamiltonian error rates (the v_i) are chosen uniformly at random from choose $[0, \sqrt{\varepsilon}]$, while the rates (the v_i) for all other errors are chosen uniformly from $[0, \varepsilon]$. We do this for 785 different error models, varying ε from $\varepsilon = .001$ to $\varepsilon = .01$. The results of the numerical experiments are shown in Figs. 9 and 10.

For each error model, Fig. 9 shows the prediction for r_Ω from the $\mathscr{L}$ matrix theory (which is r_γ) versus an estimate $\hat{r}_\Omega$ for r_Ω obtained by simulating direct RB circuits, and fitting the data to $S_d = A + Bp^d$. Error bars are 1σ, and they are computed using a standard bootstrap. We observe good agreement between r_γ (the theory's prediction) and r_Ω. However, quantifying any small underlying discrepancies between r_Ω and r_γ is difficult due to the uncertainties in the $\hat{r}_\Omega$—which arise from simulating a finite set of direct RB circuits. To remove this uncertainty, we computed the exact S_d values, using our exact $\mathscr{R}$ matrix theory for direct RB, and we then fit these exact values for S_d to $S_d = A + Bp^d$ to obtain r_Ω. Figure 10 shows r_Ω versus r_γ, and a histogram (inset) of the relative error $(r_\Omega - r_\gamma)/r_\Omega$. We find extremely close agreement between r_Ω and r_γ, with no more than $\pm 1\%$ relative error. This therefore demonstrates the accuracy of our $\mathscr{L}$ matrix theory for direct RB.

7 Conclusion

Direct RB [20] is a method for benchmarking a set of quantum gates that does not require the inflexible circuit sampling and the complex gate-compilation of standard Clifford group RB [7]. This means that direct RB can benchmark more qubits than Clifford group RB, and direct RB can reliably extract information about the performance of the native gates of a device. In this paper we have presented a general definition for the direct RB protocol, and theories that explain how and why direct RB is broadly reliable. Our theory shows that direct RB is reliable as long as, whenever an error occurs, it almost certainly spreads on to many qubits before another error occurs. This proof that direct RB is reliable does not rely on convergence of a sequence of gates to a uniformly element from a unitary 2-design. This is crucial for proving that direct RB can be reliable on $n \gg 1$ qubits with realistic error rates, as the number of layers required for such convergence necessarily increases with n. Furthermore, our theories show that the direct RB error rate can be interpreted as the average infidelity of the benchmarked gates, building on similar results for Clifford group RB [21, 22].

Clifford group RB has been extended to an entire suite of methods, including methods for estimating the error rates of individual gates [13, 91, 92], for quantifying the coherent component of gate errors [93, 94, 95], for estimating leakage and loss [96, 89, 97], for calibrating gates [40, 41], and for quantify crosstalk [85, 18]. We anticipate that all of these methods can be adapted to the more flexible direct RB framework, and fully developing these techniques is an interesting direction for future research. One limitation of direct RB is that it is not indefinitely scalable—as shown in Fig. 3—because initialization and measurement of a random stabilizer state requires $O(n^2/\log n)$ two-qubit gates. Recent work on mirror RB [43, 51, 98] has shown that it is possible to adapt direct RB to remove these expensive steps, and much of the theory and many of the analysis techniques used in this paper are also applicable to that more scalable protocol.

Acknowledgements

The authors thank Jahan Claes for providing a simplification to the proof of Proposition 2. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Quantum Testbed Pathfinder. This research was funded, in part, by the Office of the Director of National Intelligence (ODNI), Intelligence

Advanced Research Projects Activity (IARPA). Sandia National Laboratories is a multi-program laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the U.S. Department of Energy, IARPA, the ODNI, or the U.S. Government. AMP acknowledges funding from NSF award DMR-1747426 and a NASA Space Technology Graduate Research Opportunity award.

References

[1] Seth T Merkel, Jay M Gambetta, John A Smolin, Stefano Poletto, Antonio D Córcoles, Blake R Johnson, Colm A Ryan, and Matthias Steffen. "Self-consistent quantum process tomography". Physical Review A **87**, 062119 (2013). url: https://doi.org/10.1103/PhysRevA.87.062119.

[2] 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**, 14485 (2017). url: https://doi.org/10.1038/ncomms14485.

[3] Erik Nielsen, John King Gamble, Kenneth Rudinger, Travis Scholten, Kevin Young, and Robin Blume-Kohout. "Gate set tomography". Quantum **5**, 557 (2021). url: https://doi.org/10.22331/q-2021-10-05-557.

[4] Joseph Emerson, Robert Alicki, and Karol Życzkowski. "Scalable noise estimation with random unitary operators". Journal of Optics B: Quantum and Semiclassical Optics **7**, S347 (2005). url: https://doi.org/10.1088/1464-4266/7/10/021.

[5] 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). url: https://doi.org/10.1126/science.1145699.

[6] Emanuel Knill, D Leibfried, R Reichle, J Britton, RB Blakestad, JD Jost, C Langer, R Ozeri, S Seidelin, and DJ Wineland. "Randomized benchmarking of quantum gates". Physical Review A **77**, 012307 (2008). url: https://doi.org/10.1103/PhysRevA.77.012307.

[7] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. "Scalable and robust randomized benchmarking of quantum processes". Physical Review Letters **106**, 180504 (2011). url: https://doi.org/10.1103/PhysRevLett.106.180504.

[8] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. "Characterizing quantum gates via randomized benchmarking". Physical Review A **85**, 042311 (2012). url: https://doi.org/10.1103/PhysRevA.85.042311.

[9] Arnaud Carignan-Dugas, Joel J Wallman, and Joseph Emerson. "Characterizing universal gate sets via dihedral benchmarking". Physical Review A **92**, 060302 (2015). url: https://doi.org/10.1103/PhysRevA.92.060302.

[10] 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**, 16012 (2016). url: https://doi.org/10.1038/npjqi.2016.12.

[11] Winton G. Brown and Bryan Eastin. "Randomized benchmarking with restricted gate sets". Physical Review A **97**, 062323 (2018). url: https://doi.org/10.1103/PhysRevA.97.062323.

[12] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman. "Real randomized benchmarking". Quantum **2**, 85 (2018). url: https://doi.org/10.22331/q-2018-08-22-85.

[13] Easwar Magesan, Jay M Gambetta, Blake R Johnson, Colm A Ryan, Jerry M Chow, Seth T Merkel, Marcus P da Silva, George A Keefe, Mary B Rothwell, Thomas A Ohki, et al. "Efficient measurement of quantum gate error by interleaved randomized benchmarking". Physical Review Letters **109**, 080505 (2012). url: https://doi.org/10.1103/PhysRevLett.109.080505.

[14] Jonas Helsen, Xiao Xue, Lieven M. K. Vandersypen, and Stephanie Wehner. "A new class of efficient randomized benchmarking protocols". npj Quantum Information**5** (2019). url: https://doi.org/10.1038/s41534-019-0182-7.

[15] 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). url: https://doi.org/10.22331/q-2022-02-21-657.

[16] J. Helsen, I. Roth, E. Onorati, A.H. Werner, and J. Eisert. "General framework for randomized benchmarking". PRX Quantum**3** (2022). url: https://doi.org/10.1103/prxquantum.3.020357.

[17] Jahan Claes, Eleanor Rieffel, and Zhihui Wang. "Character randomized benchmarking for non-multiplicity-free groups with applications to subspace, leakage, and matchgate randomized benchmarking". PRX Quantum**2** (2021). url: https://doi.org/10.1103/prxquantum.2.010351.

[18] David C McKay, Andrew W Cross, Christopher J Wood, and Jay M Gambetta. "Correlated randomized benchmarking" (2020). arXiv:2003.02354.

[19] Jonas Helsen, Joel J. Wallman, Steven T. Flammia, and Stephanie Wehner. "Multiqubit randomized benchmarking using few samples". Physical Review A**100** (2019). url: https://doi.org/10.1103/physreva.100.032304.

[20] 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** (2019). url: https://doi.org/10.1103/PhysRevLett.123.030503.

[21] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout. "What randomized benchmarking actually measures". Physical Review Letters **119**, 130502 (2017). url: https://doi.org/10.1103/PhysRevLett.119.130502.

[22] Joel J Wallman. "Randomized benchmarking with gate-dependent noise". Quantum **2**, 47 (2018). url: https://doi.org/10.22331/q-2018-01-29-47.

[23] Maris Ozols. "Clifford group". Essays at University of Waterloo, Spring (2008).

[24] Robert Koenig and John A Smolin. "How to efficiently select an arbitrary Clifford group element". Journal of Mathematical Physics **55**, 122202 (2014). url: https://doi.org/10.1063/1.4903507.

[25] Scott Aaronson and Daniel Gottesman. "Improved simulation of stabilizer circuits". Physical Review A **70**, 052328 (2004). url: https://doi.org/10.1103/PhysRevA.70.052328.

[26] Ketan N Patel, Igor L Markov, and John P Hayes. "Optimal synthesis of linear reversible circuits". Quantum Information and Computing **8**, 282–294 (2008). url: https://doi.org/10.26421/QIC8.3-4-4.

[27] Sergey Bravyi and Dmitri Maslov. "Hadamard-free circuits expose the structure of the clifford group". IEEE Transactions on Information Theory **67**, 4546–4563 (2021). url: https://doi.org/10.1109/tit.2021.3081415.

[28] Jun Yoneda, Kenta Takeda, Tomohiro Otsuka, Takashi Nakajima, Matthieu R Delbecq, Giles Allison, Takumu Honda, Tetsuo Kodera, Shunri Oda, Yusuke Hoshi, et al. "A quantum-dot spin qubit with coherence limited by charge noise and fidelity higher than 99.9%". Nature Nanotechnology **13**, 102 (2018). url: https://doi.org/10.1038/s41565-017-0014-x.

[29] David M Zajac, Anthony J Sigillito, Maximilian Russ, Felix Borjans, Jacob M Taylor, Guido Burkard, and Jason R Petta. "Resonantly driven CNOT gate for electron spins". Science **359**, 439–442 (2018). url: https://doi.org/10.1126/science.aao5965.

[30] TF Watson, SGJ Philips, Erika Kawakami, DR Ward, Pasquale Scarlino, Menno Veldhorst, DE Savage, MG Lagally, Mark Friesen, SN Coppersmith, et al. "A programmable two-qubit quantum processor in silicon". Nature **555**, 633 (2018). url: https://doi.org/10.1038/nature25766.

[31] John M Nichol, Lucas A Orona, Shannon P Harvey, Saeed Fallahi, Geoffrey C Gardner, Michael J Manfra, and Amir Yacoby. "High-fidelity entangling gate for double-quantum-dot spin qubits". NPJ Quantum Information **3**, 3 (2017). url: https://doi.org/10.1038/s41534-016-0003-1.

[32] M Veldhorst, JCC Hwang, CH Yang, AW Leenstra, B De Ronde, JP Dehollain, JT Muhonen, FE Hudson, Kohei M Itoh, A Morello, et al. "An addressable quantum dot qubit with fault-tolerant control-fidelity". Nature Nanotechnology **9**, 981–985 (2014). url: https://doi.org/10.1038/nnano.2014.216.

[33] Antonio D Córcoles, Jay M Gambetta, Jerry M Chow, John A Smolin, Matthew Ware, Joel Strand, Britton LT Plourde, and Matthias Steffen. "Process verification of two-qubit quantum gates by randomized benchmarking". Physical Review A **87**, 030301 (2013). url: https://doi.org/10.1103/PhysRevA.87.030301.

[34] T Xia, M Lichtman, K Maller, AW Carr, MJ Piotrowicz, L Isenhower, and M Saffman. "Randomized benchmarking of single-qubit gates in a 2D array of neutral-atom qubits". Physical Review Letters **114**, 100503 (2015). url: https://doi.org/10.1103/PhysRevLett.114.100503.

[35] AD Córcoles, Easwar Magesan, Srikanth J Srinivasan, Andrew W Cross, M Steffen, Jay M Gambetta, and Jerry M Chow. "Demonstration of a quantum error detection code using a square lattice of four superconducting qubits". Nature Communications **6**, 6979 (2015). url: https://doi.org/10.1038/ncomms7979.

[36] Zijun Chen, Julian Kelly, Chris Quintana, R Barends, B Campbell, Yu Chen, B Chiaro, A Dunsworth, AG Fowler, E Lucero, et al. "Measuring and suppressing quantum state leakage in a superconducting qubit". Physical Review Letters **116**, 020501 (2016). url: https://doi.org/10.1103/PhysRevLett.116.020501.

[37] JT Muhonen, A Laucht, S Simmons, JP Dehollain, R Kalra, FE Hudson, S Freer, Kohei M Itoh, DN Jamieson, JC McCallum, et al. "Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking". Journal of Physics: Condensed Matter **27**, 154205 (2015). url: https://doi.org/10.1088/0953-8984/27/15/154205.

[38] R Barends, J Kelly, A Megrant, A Veitia, D Sank, E Jeffrey, TC White, J Mutus, AG Fowler, B Campbell, et al. "Superconducting quantum circuits at the surface code threshold for fault tolerance". Nature **508**, 500–503 (2014). url: https://doi.org/10.1038/nature13171.

[39] J Raftery, A Vrajitoarea, G Zhang, Z Leng, SJ Srinivasan, and AA Houck. "Direct digital synthesis of microwave waveforms for quantum computing" (2017). arXiv:1703.00942.

[40] MA Rol, CC Bultink, TE O'Brien, SR de Jong, LS Theis, X Fu, F Luthi, RFL Vermeulen, JC de Sterke, A Bruno, et al. "Restless tuneup of high-fidelity qubit gates". Physical Review Appl. **7**, 041001 (2017). url: https://doi.org/10.1103/PhysRevApplied.7.041001.

[41] J Kelly, R Barends, B Campbell, Y Chen, Z Chen, B Chiaro, A Dunsworth, AG Fowler, I-C Hoi, E Jeffrey, et al. "Optimal quantum control using randomized benchmarking". Physical Review Letters **112**, 240504 (2014). url: https://doi.org/10.1103/PhysRevLett.112.240504.

[42] David C McKay, Sarah Sheldon, John A Smolin, Jerry M Chow, and Jay M Gambetta. "Three qubit randomized benchmarking". Physical Review Letters **122**, 200502 (2019). url: https://doi.org/10.1103/PhysRevLett.122.200502.

[43] Timothy Proctor, Stefan Seritan, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young. "Scalable randomized benchmarking of quantum computers using mirror circuits". Physical Review Letters**129** (2022). url: https://doi.org/10.1103/physrevlett.129.150502.

[44] David Gross, Koenraad Audenaert, and Jens Eisert. "Evenly distributed unitaries: on the structure of unitary designs". Journal of Mathematical Physics **48**, 052104 (2007). url: https://doi.org/10.1063/1.2716992.

[45] 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**, 012304 (2009). url: https://doi.org/10.1103/PhysRevA.80.012304.

[46] Joel J Wallman and Steven T Flammia. "Randomized benchmarking with confidence". New Journal of Physics **16**, 103032 (2014). url: https://doi.org/10.1088/1367-2630/16/10/103032.

[47] Jeffrey M Epstein, Andrew W Cross, Easwar Magesan, and Jay M Gambetta. "Investigating the limits of randomized benchmarking protocols". Physical Review A **89**, 062321 (2014). url: https://doi.org/10.1103/PhysRevA.89.062321.

[48] D S França and A K Hashagen. "Approximate randomized benchmarking for finite groups". Journal of Physics A: Mathematical and Theoretical **51**, 395302 (2018). url: https://doi.org/10.1088/1751-8121/aad6fa.

[49] Markus Heinrich, Martin Kliesch, and Ingo Roth. "Randomized benchmarking with random quantum circuits" (2022). arXiv:2212.06181.

[50] CA Ryan, Martin Laforest, and Raymond Laflamme. "Randomized benchmarking of single- and multi-qubit control in liquid-state nmr quantum information processing". New Journal of Physics **11**, 013034 (2009). url: https://doi.org/10.1088/1367-2630/11/1/013034.

[51] Jordan Hines, Marie Lu, Ravi K Naik, Akel Hashim, Jean-Loup Ville, Brad Mitchell, John Mark Kriekebaum, David I Santiago, Stefan Seritan, Erik Nielsen, Robin Blume-Kohout, Kevin Young, Irfan Siddiqi, Birgitta Whaley, and Timothy Proctor. "Demonstrating scalable randomized benchmarking of universal gate sets". Phys. Rev. X **13**, 041030 (2023). url: https://doi.org/10.1103/PhysRevX.13.041030.

[52] Jordan Hines, Daniel Hothem, Robin Blume-Kohout, Birgitta Whaley, and Timothy Proctor. "Fully scalable randomized benchmarking without motion reversal". PRX quantum **5**, 030334 (2024). url: https://doi.org/10.1103/PRXQuantum.5.030334.

[53] Emanuel Knill. "Quantum computing with realistically noisy devices". Nature **434**, 39–44 (2005). url: https://doi.org/10.1038/nature03350.

[54] Matthew Ware, Guilhem Ribeill, Diego Ristè, Colm A. Ryan, Blake Johnson, and Marcus P. da Silva. "Experimental pauli-frame randomization on a superconducting qubit". Physical Review A**103** (2021). url: https://doi.org/10.1103/physreva.103.042604.

[55] Joel J Wallman and Joseph Emerson. "Noise tailoring for scalable quantum computation via randomized compiling". Physical Review A **94**, 052325 (2016). url: https://doi.org/10.1103/PhysRevA.94.052325.

[56] Akel Hashim, Ravi K. Naik, Alexis Morvan, Jean-Loup Ville, Bradley Mitchell, John Mark Kreikebaum, Marc Davis, Ethan Smith, Costin Iancu, Kevin P. O'Brien, Ian Hincks, Joel J. Wallman, Joseph Emerson, and Irfan Siddiqi. "Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor". Physical Review X**11** (2021). url: https://doi.org/10.1103/physrevx.11.041039.

[57] Winton Brown and Omar Fawzi. "Decoupling with random quantum circuits". Communications in Mathematical Physics **340**, 867–900 (2015). url: https://doi.org/10.1007/s00220-015-2470-1.

[58] Yasuhiro Sekino and Leonard Susskind. "Fast scramblers". Journal of High Energy Physics **2008**, 065 (2008). url: https://doi.org/10.1088/1126-6708/2008/10/065.

[59] Fernando GSL Brandão, Aram W Harrow, and Michał Horodecki. "Local random quantum circuits are approximate polynomial-designs". Communications in Mathematical Physics **346**, 397–434 (2016). url: https://doi.org/10.1007/s00220-016-2706-8.

[60] Michał Oszmaniec, Adam Sawicki, and Michał Horodecki. "Epsilon-nets, unitary designs, and random quantum circuits". IEEE Transactions on Information Theory **68**, 989–1015 (2022). url: https://doi.org/10.1109/TIT.2021.3128110.

[61] Nicholas Hunter-Jones. "Operator growth in random quantum circuits with symmetry" (2018). url: https://doi.org/10.48550/arXiv.1812.08219.

[62] C W von Keyserlingk, Tibor Rakovszky, Frank Pollmann, and S L Sondhi. "Operator hydrodynamics, OTOCs, and entanglement growth in systems without conservation laws". Physical Review X **8**, 021013 (2018). url: https://doi.org/10.1103/PhysRevX.8.021013.

[63] Adam Nahum, Sagar Vijay, and Jeongwan Haah. "Operator spreading in random unitary circuits". Physical Review X **8**, 021014 (2018). url: https://doi.org/10.1103/PhysRevX.8.021014.

[64] Seth T. Merkel, Emily J. Pritchett, and Bryan H. Fong. "Randomized benchmarking as convolution: Fourier analysis of gate dependent errors". Quantum **5**, 581 (2021). url: https://doi.org/10.22331/q-2021-11-16-581.

[65] Arnaud Carignan-Dugas, Kristine Boone, Joel J Wallman, and Joseph Emerson. "From randomized benchmarking experiments to gate-set circuit fidelity: how to interpret randomized benchmarking decay parameters". New Journal of Physics **20**, 092001 (2018). url: https://doi.org/10.1088/1367-2630/aadcc7.

[66] Michael A Nielsen. "A simple formula for the average gate fidelity of a quantum dynamical operation". Physics Letters A **303**, 249–252 (2002). url: https://doi.org/10.1016/S0375-9601(02)01272-0.

[67] S Mavadia, CL Edmunds, C Hempel, H Ball, F Roy, TM Stace, and MJ Biercuk. "Experimental quantum verification in the presence of temporally correlated noise". npj Quantum Information **4**, 7 (2018). url: https://doi.org/10.1038/s41534-017-0052-0.

[68] MA Fogarty, M Veldhorst, R Harper, CH Yang, SD Bartlett, ST Flammia, and AS Dzurak. "Nonexponential fidelity decay in randomized benchmarking with low-frequency noise". Physical Review A **92**, 022326 (2015). url: https://doi.org/10.1103/PhysRevA.92.022326.

[69] Bryan H Fong and Seth T Merkel. "Randomized benchmarking, correlated noise, and ising models" (2017). arXiv:1703.09747.

[70] Timothy Proctor, Melissa Revelle, Erik Nielsen, Kenneth Rudinger, Daniel Lobser, Peter Maunz, Robin Blume-Kohout, and Kevin Young. "Detecting and tracking drift in quantum information processors". Nature Communications **11**, 5396 (2020). url: https://doi.org/10.1038/s41467-020-19074-4.

[71] Christopher Granade, Christopher Ferrie, and DG Cory. "Accelerated randomized benchmarking". New Journal of Physics **17**, 013042 (2015). url: https://doi.org/10.1088/1367-2630/17/1/013042.

[72] Ian Hincks, Joel J. Wallman, Chris Ferrie, Chris Granade, and David G. Cory. "Bayesian inference for randomized benchmarking protocols." (2018). arXiv:1802.00401.

[73] G Vidal and C M Dawson. "Universal quantum circuit for two-qubit transformations with three controlled-NOT gates". Physical Review A **69**, 010301 (2004). url: https://doi.org/10.1103/PhysRevA.69.010301.

[74] Vivek V Shende, Stephen S Bullock, and Igor L Markov. "Recognizing small-circuit structure in two-qubit operators". Physical Review A **70**, 012310 (2004). url: https://doi.org/10.1103/PhysRevA.70.012310.

[75] Shelly Garion, Naoki Kanazawa, Haggai Landa, David C. McKay, Sarah Sheldon, Andrew W. Cross, and Christopher J. Wood. "Experimental implementation of non-Clifford interleaved randomized benchmarking with a controlled-s gate". Physical Review Research**3** (2021). url: https://doi.org/10.1103/physrevresearch.3.013204.

[76] 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**, 505–510 (2019). url: https://doi.org/10.1038/s41586-019-1666-5.

[77] Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven. "Characterizing quantum supremacy in near-term devices". Nature Physics **14**, 595 (2018). url: https://doi.org/10.1038/s41567-018-0124-x.

[78] Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout. "Measuring the capabilities of quantum computers". Nature Physics **18**, 75–79 (2021). url: https://doi.org/10.1038/s41567-021-01409-7.

[79] Erik Nielsen, Robin Blume-Kohout, Lucas Saldyt, Jonathan Gross, Travis Scholten, Kenneth Rudinger, Timothy Proctor, and John King Gamble. "Pygsti version 0.9.6" (2018).

[80] Erik Nielsen, Kenneth Rudinger, Timothy Proctor, Antonio Russo, Kevin Young, and Robin Blume-Kohout. "Probing quantum processor performance with pyGSTi". Quantum Sci. Technol. **5**, 044002 (2020). url: https://doi.org/10.1088/2058-9565/ab8aa4.

[81] Robin Harper, Ian Hincks, Chris Ferrie, Steven T Flammia, and Joel J Wallman. "Statistical analysis of randomized benchmarking". Physical Review A **99**, 052350 (2019). url: https://doi.org/10.1103/PhysRevA.99.052350.

[82] 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". Nature Communications **10**, 5347 (2019). url: https://doi.org/10.1038/s41467-019-13068-7.

[83] Kristine Boone, Arnaud Carignan-Dugas, Joel J. Wallman, and Joseph Emerson. "Randomized benchmarking under different gate sets". Physical Review A**99** (2019). url: https://doi.org/10.1103/physreva.99.032329.

[84] Steven T Flammia and Yi-Kai Liu. "Direct fidelity estimation from few Pauli measurements". Phys. Rev. Lett. **106**, 230501 (2011). url: http://doi.org/10.1103/PhysRevLett.106.230501.

[85] Jay M Gambetta, AD Córcoles, Seth T Merkel, Blake R Johnson, John A Smolin, Jerry M Chow, Colm A Ryan, Chad Rigetti, S Poletto, Thomas A Ohki, et al. "Characterization of addressability by simultaneous randomized benchmarking". Physical Review Letters **109**, 240504 (2012). url: https://doi.org/10.1103/PhysRevLett.109.240504.

[86] Persi Diaconis. "Group representations in probability and statistics". Volume 11 of Lecture Notes-Monograph Series. Institute of Mathematical Statistics. Hayways, CA (1988). url: https://doi.org/10.1214/lnms/1215467410.

[87] Persi Diaconis and Laurent Saloff-Coste. "Random walks on finite groups: a survey of analytic techniques". Probability measures on groups and related structures, XI (Oberwolfach, 1994)Pages 44–75 (1995). url: https://doi.org/10.1142/2760.

[88] Laurent Saloff-Coste. "Random walks on finite groups". In Probability on Discrete Structures. Pages 263–346. Springer Berlin Heidelberg (2004). url: https://doi.org/10.1007/978-3-662-09444-0_5.

[89] T Chasseur and FK Wilhelm. "Complete randomized benchmarking protocol accounting for leakage errors". Physical Review A **92**, 042333 (2015). url: https://doi.org/10.1103/PhysRevA.92.042333.

[90] Robin Blume-Kohout, Marcus P. da Silva, Erik Nielsen, Timothy Proctor, Kenneth Rudinger, Mohan Sarovar, and Kevin Young. "A taxonomy of small Markovian errors". PRX Quantum**3** (2022). url: https://doi.org/10.1103/prxquantum.3.020335.

[91] Robin Harper and Steven T Flammia. "Estimating the fidelity of T gates using standard interleaved randomized benchmarking". Quantum Science and Technology **2**, 015008 (2017). url: https://doi.org/10.1088/2058-9565/aa5f8d.

[92] Tobias Chasseur, Daniel M Reich, Christiane P Koch, and Frank K Wilhelm. "Hybrid benchmarking of arbitrary quantum gates". Physical Review A **95**, 062335 (2017). url: https://doi.org/10.1103/PhysRevA.95.062335.

[93] Joel Wallman, Chris Granade, Robin Harper, and Steven T Flammia. "Estimating the coherence of noise". New Journal of Physics **17**, 113020 (2015). url: https://doi.org/10.1088/1367-2630/17/11/113020.

[94] Guanru Feng, Joel J Wallman, Brandon Buonacorsi, Franklin H Cho, Daniel K Park, Tao Xin, Dawei Lu, Jonathan Baugh, and Raymond Laflamme. "Estimating the coherence of noise in quantum control of a solid-state qubit". Physical Review Letters **117**, 260501 (2016). url: https://doi.org/10.1103/PhysRevLett.117.260501.

[95] Sarah Sheldon, Lev S Bishop, Easwar Magesan, Stefan Filipp, Jerry M Chow, and Jay M Gambetta. "Characterizing errors on qubit operations via iterative randomized benchmarking". Physical Review A **93**, 012301 (2016). url: https://doi.org/10.1103/PhysRevA.93.012301.

[96] Christopher J Wood and Jay M Gambetta. "Quantification and characterization of leakage errors". Physical Review A **97**, 032306 (2018). url: https://doi.org/10.1103/PhysRevA.97.032306.

[97] Joel J Wallman, Marie Barnhill, and Joseph Emerson. "Robust characterization of loss rates". Physical Review Letters **115**, 060501 (2015). url: https://doi.org/10.1103/PhysRevLett.115.060501.

[98] Karl Mayer, Alex Hall, Thomas Gatterman, Si Khadir Halit, Kenny Lee, Justin Bohnet, Dan Gresh, Aaron Hankin, Kevin Gilmore, and John Gaebler. "Theory of mirror benchmarking and demonstration on a quantum computer" (2021). url: https://doi.org/10.48550/arXiv.2108.10431.

A Proof of Proposition 3

Proof. By linearity $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E}_1 + \mathcal{E}_\gamma) = \mathcal{E}_1 + \gamma\mathcal{E}_\gamma$. If $\gamma = 1$ then the proposition holds trivially, so consider the case of $\gamma < 1$. If $\gamma < 1$ then, because $\tilde{\mathbb{G}}$ is a low-error implementation of $\mathbb{G}$, both 1 and γ are non-degenerate eigenvalues of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$. Therefore, up to constant scalings, $\mathcal{E}_1$ and $\mathcal{E}_\gamma$ are unique. Define

$$\tilde{\mathcal{G}}_{\Omega-\text{avg}} \equiv \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})\tilde{\mathcal{G}}, \quad \mathcal{G}_{\Omega-\text{avg}} \equiv \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})\mathcal{G}, \tag{117}$$

where we abuse notation and consider all gates as these as superoperators. Let $\langle\!\langle V|$ denote some left eigenvector of the Hilbert-Schmidt space matrix $\tilde{\mathcal{G}}_{\Omega-\text{ avg}}$ with eigenvalue 1, i.e., $\langle\!\langle V|\tilde{\mathcal{G}}_{\Omega-\text{ avg}} = \langle\!\langle V|$. This eigenvector exists because the elements of $\tilde{\mathbb{G}}$ are trace-preserving maps, which implies that $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$ is also a trace-preserving map, and all trace-preserving maps have 1 as an eigenvalue. Because the the elements of $\mathbb{G}$ are the superoperators for unitary gates, $\mathcal{G}^{-1}|B_0\rangle\!\rangle = |B_0\rangle\!\rangle$ for all $\mathcal{G} \in \mathbb{G}$, where $B_0 = \mathbb{1}_d/\sqrt{d}$ (here we are using the basis for Hilbert-Schmidt space introduced in Section 5.1). This is the statement that unitary maps are unital. Hence we have that

$$\begin{aligned}\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(|B_0\rangle\!\rangle\langle\!\langle V|) &= \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})[\mathcal{G}^{-1}|B_0\rangle\!\rangle\langle\!\langle V|\tilde{\mathcal{G}}] \\ &= |B_0\rangle\!\rangle\langle\!\langle V|\tilde{\mathcal{G}}_{\Omega-\text{avg}}, \\ &= |B_0\rangle\!\rangle\langle\!\langle V|\end{aligned} \tag{118}$$

and therefore

$$\mathcal{E}_1 = |B_0\rangle\!\rangle\langle\!\langle V|. \tag{119}$$

Using the "stacked" representation discussed in Section 5.1, we may turn $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E}_\gamma) = \gamma\mathcal{E}_\gamma$ into a more standard eigenvector equation.

Let $\mathcal{G}_{\text{u}}^{-1} = \mathcal{G}^{-1} - |B_0\rangle\!\rangle\langle\!\langle B_0|$ for all $\mathcal{G} \in \mathbb{G}$, which is known as the unital component of $\mathcal{G}^{-1}$. Because the elements in $\mathbb{G}$ are unitary gates

$$\langle\!\langle B_0|\mathcal{G}_{\text{u}}^{-1} = \mathcal{G}_{\text{u}}^{-1}|B_0\rangle\!\rangle = 0, \tag{120}$$

for all $\mathcal{G} \in \mathbb{G}$. Therefore, for any superoperator $\mathcal{X}$,

$$\begin{aligned}\text{vec}(\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{X})) &= \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})[\tilde{\mathcal{G}}^T \otimes \mathcal{G}^{-1}]\,\text{vec}(\mathcal{X}) \\ &= \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})[\tilde{\mathcal{G}}^T \otimes (|B_0\rangle\!\rangle\langle\!\langle B_0| + \mathcal{G}_{\text{u}}^{-1})]\,\text{vec}(\mathcal{X}) \\ &= [\mathcal{L}_0 + \mathcal{L}_\perp]\,\text{vec}(\mathcal{X})\end{aligned} \tag{121}$$

with $\mathcal{L}_0 = \tilde{\mathcal{G}}_{\Omega-\text{avg}}^T \otimes |B_0\rangle\!\rangle\langle\!\langle B_0|$ and

$$\mathcal{L}_\perp = \sum_{\mathcal{G}\in\mathbb{G}} \Omega(\mathcal{G})(\tilde{\mathcal{G}}^T \otimes \mathcal{G}_{\text{u}}^{-1}). \tag{122}$$

Because $\mathcal{G}_{\text{u}}^{-1}|B_0\rangle\!\rangle\langle\!\langle B_0| = |B_0\rangle\!\rangle\langle\!\langle B_0|\mathcal{G}_{\text{u}}^{-1} = 0$, we have that the $\mathcal{L}_0$ and $\mathcal{L}_\perp$ matrices satisfy

$$\mathcal{L}_0\mathcal{L}_\perp = \mathcal{L}_\perp\mathcal{L}_0 = 0. \tag{123}$$

As such, the set of eigenvalues of $\mathscr{L}$ is the union of the sets of eigenvalues of $\mathcal{L}_0$ and $\mathcal{L}_\perp$. Moreover, if $\mathcal{V}$ is an eigen-operator of $\mathscr{L}$ and if the associated eigenvalue is not an eigenvalue of $\mathcal{L}_0$, then

$$\text{vec}(\mathcal{V}) = \sum_j \sum_{k>0} v_{jk}|B_j\rangle\!\rangle \otimes |B_k\rangle\!\rangle, \tag{124}$$

for some v_{jk}, and so $\langle\!\langle B_0|\mathcal{V} = 0$.

We now show that γ is not an eigenvalue of $\mathcal{L}_0$, which then implies that $\langle\!\langle B_0|\mathcal{E}_\gamma = 0$. The eigenvalues of $\mathcal{L}_0$ are the eigenvalues of $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$, together with 0. $\mathcal{G}_{\Omega-\text{avg}}$ is a sequence-asymptotic unitary 2-design, and thus has 1 as a non-degenerate eigenvalue. We will show this is also true for $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$.

In [22], the author shows that, if the generating set is a 2-design, $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$ will generally have a unique largest eigenvalue. We extend this argument to sequence-asymptotic unitary 2-designs. The limiting map of the generator twirl is a 2-design twirl, and so in the limit the result from [22] holds. Now, suppose that it did not hold that the generator twirl had a non-degenerate 1-eigenspace. At no point in the limiting sequence could it acquire one, thus it has to have one to begin with.

Now, suppose by way of contradiction that γ is an eigenvalue of $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$. Since γ is a perturbation of 1, in the limit that the perturbation goes to 0, $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$ would acquire 1 as a degenerate eigenvalue, thus γ cannot be an eigenvalue of $\tilde{\mathcal{G}}_{\Omega-\text{avg}}$.

Thus, $\mathcal{E}_\gamma$ is not an eigenvector of $\mathcal{L}_0$ and hence

$$\langle\!\langle B_0|\mathcal{E}_\gamma = 0. \tag{125}$$

This equality implies that

$$\mathcal{E}_\gamma = \sum_i \sum_{j>0} \alpha_{ij} |B_j\rangle\!\rangle\langle\!\langle B_i|, \tag{126}$$

for some α_{ij}. A depolarization channel $\mathcal{D}_\lambda$ maps $\rho \to \rho$ with probability λ and $\rho \to \mathbb{1}/d$ with probability $1-\lambda$. As such, in Hilbert-Schmidt space with the basis considered here

$$\mathcal{D}_\lambda = |B_0\rangle\!\rangle\langle\!\langle B_0| + \lambda \sum_{i\neq 0} |B_i\rangle\!\rangle\langle\!\langle B_i|. \tag{127}$$

Hence, from Eqs. (126) and (119) we have that $\mathcal{E}_1 + \gamma\mathcal{E}_\gamma = \mathcal{D}_\gamma(\mathcal{E}_1 + \mathcal{E}_\gamma)$. Finally, using this relation and considering the action of $\mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}$ on $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_\gamma$, we have that

$$\begin{aligned} \mathscr{L}_{\mathbb{G},\tilde{\mathbb{G}},\Omega}(\mathcal{E}) &= \mathcal{E}_1 + \gamma\mathcal{E}_\gamma, \\ &= \mathcal{D}_\gamma(\mathcal{E}_1 + \mathcal{E}_\gamma), \\ &= \mathcal{D}_\gamma\mathcal{E}, \end{aligned} \tag{128}$$

which concludes the proof. □

Note that the $\mathcal{E}$ in this proposition is not necessarily a completely positive map, and generically it is not.

B Sampling distributions

In this appendix we present one possible family of sampling distributions (Ω), called the *edge grab* sampler and first introduced in [78]. This sampling distribution is designed for sampling from an n-qubit gate set $\mathbb{G}$ containing gates consisting of a single layer of parallel one- and two-qubit gates, with one-qubit gates from some set $\mathbb{G}_1$ and two-qubit gates only between a connected qubits as specified by an edge list E (that can contain directed edges). The sampling distribution is parameterized by the mean two-qubit gate of the sampled gates $\bar{\xi}$. The two-qubit gate density of an n-qubit layer is simply $\xi = {}^{2\alpha}/_{n}$ where α is the number of two-qubit gates in the layer. The edge grab sampling distribution $\Omega_{\bar{\xi}}$ is most easily described by an algorithm for drawing a sample from $\Omega_{\bar{\xi}}$. Drawing a sampling consists of the following procedure:

1. *Select a candidate set of edges* E_c. Initialize E_c to the empty set, and initialize E_r to the set of all edges E. Then, until E_r is the empty set:

   1.1 Select an edge v uniformly at random from E_r.

   1.2 Add v to E_c and remove all edges that have a qubit in common with v from E_r.

2. *Select a subset of the candidate edges*. For each edge in E_c, include it in a final edge set E_f with a probability of $n\bar{\xi}/(2|E_c|)$ where $|E_c|$ is the total number of edges in E_c. If $n\bar{\xi}/(2|E_c|) > 1$ construct a new edge set sample E_c.

3. *Construct the sampled gate*. The sampled gate $G \in \mathbb{G}$ is then constructed by adding a two-qubit gate on each edge in E_f and for all remaining qubits independently and uniformly sample a single-qubit gate from $\mathbb{G}_1$ to include in G.

The expected number of two-qubit gates in G is $n\bar{\xi}/2$, so this sampler generates an n-qubit layer with an expected two-qubit gate density of $\bar{\xi}$. One useful property of this sampling is that the probability of sampling any particular n-qubit gate G is non-zero for every $G \in \mathbb{G}$ (except if $\bar{\xi} = 0$ or $\bar{\xi} = 1$). Finally, note that the edge grab algorithm is invalid if $n\bar{\xi}/(2|E_c|) > 1$ for any possible candidate edge set E_c. For an even number of fully-connected qubits, $\bar{\xi}$ can take any value between 0 and 1. But for any other connectivity the maximum achievable value of $\bar{\xi}$ is smaller. In our experiments, we set $\bar{\xi} = 1/8$. This is an achievable value of $\bar{\xi}$ in the edge grab algorithm for all the processors that we benchmarked.