Randomized benchmarking with restricted gate sets

Winton G. Brown[*] and Bryan Eastin

Northrop Grumman Corporation, Falls Church, Virginia 22042, USA

Standard randomized benchmarking protocols entail sampling from a unitary 2-design, which is not always practical. In this article we examine randomized benchmarking protocols based on subgroups of the Clifford group that are not unitary 2-designs. We introduce a general method for analyzing such protocols and subsequently apply it to two subgroups, the group generated by controlled-NOT, Hadamard, and Pauli gates and that generated by only controlled-NOT and Pauli gates. In both cases the error probability can be estimated to within a factor of 2 or less where the factor can be arranged to be conservative and to decay exponentially in the number of qubits. For randomized benchmarking of logical qubits even better accuracy will typically be obtained. Thus, we show that sampling a distribution which is close to a unitary 2-design, although sufficient, is not necessary for randomized benchmarking to high accuracy.

I. INTRODUCTION

Randomized benchmarking is widely used for characterizing the performance of quantum information processing devices [1–8]. Standard randomized benchmarking protocols require the capability to sample from a unitary 2-design [9–12], for which the Clifford group, or certain subgroups of the Clifford group which are nonetheless unitary 2-designs, are usually used. Sampling from a 2-design can be challenging, however, particularly for the case of randomized benchmarking of logical qubits, where the set of high-fidelity logical gates that can be implemented in a straightforward manner is invariably restricted.

For many quantum codes, techniques such as state injection and distillation are required to implement some generators of the logical Clifford group. Clifford gates incorporating such generators suffer from much higher overhead and probability of error. Logical qubits are likely to be in short supply for some time, so overhead is a significant concern for near-term demonstrations of randomized benchmarking on logical qubits. Furthermore, incorporating logical Clifford gates with low utility and poor performance in logical randomized benchmarking is undesirable as it results in an overly pessimistic assessment of the logical gate set. The latter issue might be resolved by benchmarking individual logical gates via interleaved randomized benchmarking [13] except that this technique has been found to perform poorly for gate sets with wildly varying error probabilities [14]. Faced with these issues, an obvious question to ask is whether the offending gates are truly necessary for randomized benchmarking. We show here that in many cases they are not.

It should be said that the question of randomized benchmarking using groups that do not form a unitary 2-design has been considered previously. Specifically, dihedral benchmarking, which utilizes the group generated by X, CX, and single-qubit phase rotations, has been proposed for use in benchmarking non-Clifford gates [15,16].

The fundamentals of randomized benchmarking are briefly reviewed in Sec. II. Section III introduces randomized benchmarking using subgroups of the Clifford group and analyzes the performance in the aforementioned cases of interest. Concluding remarks appear in Sec. IV.

II. RANDOMIZED BENCHMARKING

A. Setting and assumptions

The goal of randomized benchmarking is to determine the average fidelity of the error channels of a group of unitary gates of interest G. For the benchmarking procedure to accurately return this value, the set of error channels associated with the gates must satisfy the following properties:

(i) Markovianity. For any target $U \in$ G the true evolution of the quantum device can be expressed as a completely positive map $\mathcal{E}_U \circ \hat{U}$, where

$$\mathcal{E}_U \circ \hat{U}(\rho) = \sum_i E_i U \rho U^\dagger E_i^\dagger.$$

(ii) Weak gate dependence. The error channel does not depend strongly on the gate applied. That is, if $\mathcal{E}_U$ is the error channel associated with U and $\mathcal{E} = \frac{1}{|\mathrm{G}|} \sum_{U \in \mathrm{G}} \mathcal{E}_U$ then

$$\|\mathcal{E} - \mathcal{E}_U\|_1 \ll 1,$$

where the norm used is the induced operator 1-norm.

Recently, some complications with regard to weak gate dependence have been noted [17,18]. For simplicity, we further assume that the error channel is completely independent of the gate applied.

B. Twirling

The basic approach utilized by randomized benchmarking is to symmetrize a quantum channel $\mathcal{E}$, with respect to a group of unitary gates G, by performing a random gate from G on

─────────
[*]winton.brown@ngc.com

the input and its inverse on the output of the channel. Once symmetrized, the channel can be described by a small number of parameters which may then be determined efficiently by a simplified tomography procedure.

Specifically, performing a random gate U, sampled from G, on the input of the channel $\mathcal{E}$ while performing its inverse $U^\dagger$ on the output results in the channel

$$\tilde{\mathcal{E}}_{\mathrm{G}} = \mathbb{E}_{U \sim \mathrm{G}}[\hat{U}^\dagger \circ \mathcal{E} \circ \hat{U}],$$

where $\mathbb{E}_{U \sim \mathrm{G}}$ represents an average over all U sampled uniformly from G. The channel $\tilde{\mathcal{E}}_{\mathrm{G}}$ is symmetrized in that it is invariant under conjugation by elements of the group G, that is,

$$\hat{U}^\dagger \circ \tilde{\mathcal{E}}_{\mathrm{G}} \circ \hat{U} = \tilde{\mathcal{E}}_{\mathrm{G}} \,\forall\, U \in \mathrm{G}.$$

This symmetrization procedure is referred to as G twirling. Formally, G twirling projects the channel onto the invariants of G.

For the group $\mathrm{SU}(2^n)$ there are two linearly independent invariants under twirling. These may be taken to be the identity channel $\hat{I}(\rho) = \rho$ and the completely depolarizing channel $\mathcal{D}(\rho) = \frac{1}{4^n} \sum_\mu P_\mu \rho P_\mu$. Given an expression for an error channel in terms of Pauli operators,

$$\mathcal{E}(\rho) = \sum_{\mu\nu} x_{\mu\nu} P_\mu \rho P_\nu,$$

the corresponding fully twirled channel is

$$\tilde{\mathcal{E}}_{\mathrm{SU}(2^n)}(\rho) = (1-p)\rho + \frac{p}{4^n - 1} \sum_{\mu \neq I} P_\mu \rho P_\mu$$

$$= \lambda\rho + \frac{1-\lambda}{4^n} \mathcal{D}(\rho),$$

where $p = \sum_{\mu \neq I} x_{\mu\mu}$ is the entanglement infidelity and

$$\lambda = 1 - p\frac{4^n}{4^n - 1}$$

is the eigenvalue of the channel for nonidentity Pauli operators, i.e., $\tilde{\mathcal{E}}_{\mathrm{SU}(2^n)}(P_\mu) = \lambda P_\mu$ for all $P_\mu \neq I$. Thus, twirling by $\mathrm{SU}(2^n)$ is sufficient to convert any error channel to one specified by a single real number.

Twirling by $\mathrm{SU}(2^n)$ is not practical experimentally, but neither is it necessary. It suffices to twirl by any group G satisfying the following condition: For any polynomial that is second order and homogeneous in both the matrix elements of U and $U^\dagger$ (in the fundamental representation), the expectation value is the same whether U is sampled uniformly from $\mathrm{SU}(2^n)$ or G. Groups with this property are referred to as 2-designs [19]. A 2-design for $\mathrm{SU}(2^n)$ can equivalently be defined as a group G satisfying

$$\overline{a_{\mu\nu}(U)} = 0, \quad \overline{[a_{\mu\nu}(U)]^2} = \frac{1}{4^n - 1}$$

for all $\mu, \nu \neq I$ where the overline denotes an average over U and

$$U P_\mu U^\dagger = \sum_\nu a_{\mu\nu}(U) P_\nu.$$

In words this means that each nonidentity Pauli operator is mapped to every nonidentity Pauli operator with an average

amplitude of zero and equal average square amplitude. It is straightforward to see that the Clifford group satisfies these conditions, as well as certain subgroups thereof, such as, in the case of a single qubit, the group generated by $e^{i\pi/(3\sqrt{3})(X+Y+Z)}$ and the X gate. Twirling by 2-designs such as these forms the basis of standard randomized benchmarking.

C. Standard randomized benchmarking protocol

Given the gate error channel

$$\mathcal{E}(\rho) = \sum_{\mu\nu} x_{\mu\nu} P_\mu \rho P_\nu,$$

the goal of randomized benchmarking is to determine the entanglement infidelity

$$p = \sum_{\mu \neq I} x_{\mu\mu}$$

or some simple function thereof.[1] The standard randomized benchmarking protocol [11] consists of many repetitions of the following experiment. (i) Prepare an initial state ρ_0 such that ρ_0 is the $+1$ eigenstate of a projector $\mathbb{P}$. (ii) Perform a random gate sequence $U_{l-1}, \ldots, U_1$ where each gate is selected independently from a unitary 2-design. (iii) Perform one final unitary gate chosen such that in the absence of errors the unitary gate sequence performs the identity[2]

$$U_l = \prod_{t=1}^{l-1} U_t^{-1}.$$

(iv) Measure $\mathbb{P}$.

Averaged over many runs, the measurement statistics yield a fidelity for the experiment which depends only on the twirled error channel and the preparation and measurement errors. In order to isolate state preparation and measurement errors $\mathcal{E}_p$ and $\mathcal{E}_m$, respectively, the length l of the gate sequences is varied. The average fidelity for gate sequences of length l is given by

$$f_l = \mathrm{tr}[\mathbb{P}\mathcal{E}_m(\tilde{\mathcal{E}}^l(\mathcal{E}_p(\rho_0)))] = \mathbb{E}_{|\mathbf{U}|=l}(f_\mathbf{U}),$$

where $f_\mathbf{U}$ is the fidelity for a given sequence of unitaries $\mathbf{U} = (U_1, \ldots, U_l)$. The average sequence fidelities as a function of length are then fit to the decay curve

$$f_l = c_0 + c_1 \lambda^l,$$

where λ is as defined above.

The total number of measurements required is minimized when a different random sequence is selected for each run, but for practical reasons, each gate sequence is typically repeated many times in order to determine the fidelity $f_\mathbf{U}$ for each gate sequence $\mathbf{U}$. This also allows additional information about

---

[1]In fact, randomized benchmarking papers often quote the average infidelity, which is related to the entanglement infidelity by $p_{\mathrm{ave}} = 2^n p/(2^n + 1)$ [20]. We focus on the entanglement infidelity instead since it is more relevant for multiqubit states and corresponds to the error parameter typically used in simulations of quantum error correction.

[2]It is worth mentioning that in practice it is better to return to the starting eigenbasis rather than the starting state.

the error channel to be extracted, notably its coherence and nonunitality [9]. The minimum number of gate sequences needed to estimate f_l to a given accuracy depends on the variance of the gate sequence fidelities

$$v_l = \sum_{|\mathbf{U}|=l} (f_{\mathbf{U}} - f_l)^2,$$

for which general bounds are derived in the literature [21,22].

III. CLIFFORD SUBGROUP TWIRLING

In this section, the standard randomized benchmarking protocol is adapted to sampling from subgroups of the Clifford group which are not unitary 2-designs. All of the groups we consider contain the Pauli group, that is, the group generated by all single-qubit Pauli gates, and so are unitary 1-designs. It is important to note that twirling with respect to the Pauli group converts an arbitrary channel

$$\mathcal{E}(\rho) = \sum_{\mu\nu} x_{\mu\nu} P_\mu \rho P_\nu$$

into the corresponding stochastic Pauli channel

$$\tilde{\mathcal{E}}_{\mathrm{P}}(\rho) = \sum_\mu x_{\mu\mu} P_\mu \rho P_\mu.$$

Each Pauli operator is an eigenoperator of every Pauli channel, that is, $\tilde{\mathcal{E}}_{\mathrm{P}}(P_\mu) = \lambda_\mu P_\mu$. The corresponding eigenvalue is given by

$$\lambda_\mu = \sum_{\nu | [P_\nu, P_\mu]=0} x_{\nu\nu} - \sum_{\nu | [P_\nu, P_\mu]\neq 0} x_{\nu\nu}.$$

For any subgroup of the Clifford group S, the orbit of each Pauli operator under the action of S forms one of a set of k blocks $\{\mathbf{B}_0, \ldots, \mathbf{B}_{k-1}\}$, each containing $N_i(n)$ Pauli operators. Twirling with respect to a subgroup of the Clifford group which also contains the Pauli group therefore results in a channel of the form

$$\tilde{\mathcal{E}}_{\mathrm{S}}(\rho) = (1-p)\rho + \sum_{i=1}^{k-1} \frac{p_i}{N_i(n)} \sum_{\mu \in \mathbf{B_i}} P_\mu \rho P_\mu,$$

where $\mathbf{B}_0 = \{I\}$,

$$p_i = \sum_{\mu \in \mathbf{B_i}} x_{\mu\mu}, \quad p = \sum_{i=1}^{k-1} p_i.$$

In a case of imperfect mixing such as this, the fidelity decay curve has the form

$$f_l = c_0 + \sum_{i=1}^{k-1} c_i \lambda_i^l,$$

where

$$c_i = \mathrm{tr}\left[\mathbb{P}\mathcal{E}_m \left(\sum_{\nu \in \mathbf{B_i}} \mathrm{tr}[P_\nu \mathcal{E}_p(\rho_0)] P_\nu \right) \right],$$

ρ_0 is the ideal initial state, and $\mathbb{P}$ is the projector of interest. Notably, in the multiqubit case, the final multiqubit measurement is typically implemented via many single-qubit projective

measurements. Using the measurement results from each qubit separately, many such curves can be extracted concurrently or only a single one of particular interest (e.g., such that only one $c_{i\neq 0}$ is significant).

In theory, the parameters p_i can be determined if it is possible to prepare and measure states in the eigenspace of at least one Pauli operator from each block, either sequentially or all at once using a multiqubit initial state of the form

$$\rho_0 = \frac{1}{2^n} \prod_\mu (I + P_\mu).$$

Often, however, preparing an eigenstate for one Pauli operator from each block is impractical in situations where implementing the gate(s) required to convert between blocks is impractical.

Alternatively, if there exists a Pauli operator P that commutes with approximately the same fraction of the Pauli operators within each block, then the entanglement fidelity of the error channel can be determined approximately by preparing an eigenstate of P and measuring the decay of the expectation value of the corresponding projector as a function of gate sequence length. This is the primary approach taken in the remainder of the paper.

A. The real Clifford group

Consider the subgroup of the Clifford group that preserves the evenness or oddness of the number of Y elements in a Pauli string. This group is referred to as the real Clifford group and is generated by Hadamard, controlled-NOT, and single-qubit Pauli gates.

Twirling with respect to the real Clifford group results in a channel with two nontrivial blocks (i) $\mathbf{B}_1$ consists of the nonidentity Pauli operators with an even number of Y elements (real Pauli operators). The size of $\mathbf{B}_1$ is

$$N_1(n) = \sum_{l \in \mathrm{even}} \binom{n}{l} 3^{n-l} = \frac{4^n + 2^n}{2} - 1.$$

(ii) $\mathbf{B}_2$ consists of the Pauli operators with an odd number of Y elements (imaginary Pauli operators). The size of $\mathbf{B}_2$ is

$$N_2(n) = \sum_{l \in \mathrm{odd}} \binom{n}{l} 3^{n-l} = \frac{4^n - 2^n}{2}.$$

Given a channel

$$\mathcal{E}(\rho) = \sum_{\nu\mu} x_{\mu\nu} P_\mu \rho P_\nu,$$

twirling with respect to the real Clifford group results in the channel

$$\tilde{\mathcal{E}}_{\mathrm{R}}(\rho) = (1-p)\rho + \frac{p_1}{N_1(n)} \sum_{\mu \in \mathbf{B}_1} P_\mu \rho P_\mu$$

$$+ \frac{p_2}{N_2(n)} \sum_{\mu \in \mathbf{B}_2} P_\mu \rho P_\mu,$$

where

$$p_1 = \sum_{\mu \in \mathbf{B}_1} x_{\mu\mu}, \quad p_2 = \sum_{\mu \in \mathbf{B}_2} x_{\mu\mu}.$$

A real Pauli operator anticommutes with $N_1(n-1) + N_2(n-1) + 1$ real Pauli operators and exactly the same number of imaginary Pauli operators. An imaginary Pauli operator anticommutes with $2N_1(n-1) + 2$ real Pauli operators and $2N_2(n-1)$ imaginary Pauli operators. It follows that each real Pauli operator is an eigenvector of the twirled channel $\tilde{\mathcal{E}}_R$ with eigenvalue

$$\begin{aligned}\lambda_1 &= 1 - 2p_1 \frac{N_1(n-1) + N_2(n-1) + 1}{N_1(n)} \\ &\quad - 2p_2 \frac{N_1(n-1) + N_2(n-1) + 1}{N_2(n)} \\ &= 1 - p_1 \frac{4^n}{4^n + 2^n - 2} - p_2 \frac{2^n}{2^n - 1} \\ &= 1 - p_2 - p_1 + \frac{p_1 - p_2}{2^n} + O(2^{-2n}),\end{aligned}$$

while each imaginary Pauli operator is an eigenvector of $\tilde{\mathcal{E}}_R$ with eigenvalue

$$\begin{aligned}\lambda_2 &= 1 - 4p_1 \frac{N_1(n-1) + 1}{N_1(n)} - 4p_2 \frac{N_2(n-1)}{N_2(n)} \\ &= 1 - p_1 \frac{2^n}{2^n - 1} - p_2 \frac{4^n - 2^{n+1}}{4^n - 2^n} \\ &= 1 - p_2 - p_1 + \frac{p_2 - p_1}{2^n} + O(2^{-2n}).\end{aligned}$$

To determine the parameters p_1 and p_2 it is sufficient to prepare and measure state(s) which are eigenstates of both a real and an imaginary Pauli operator,[3] but this is typically challenging in cases where only real Clifford gates are available. In particular, codes for which only the real Clifford gates are transversal generally lack a straightforward procedure for preparation and measurement in the logical Y basis. Instead, consider the case where the initial state is only an eigenstate of real Pauli operators and therefore only λ_1 can be extracted.

Recalling that $p = p_1 + p_2$, we see that given λ_1, the entanglement infidelity can be bounded as follows:

$$\frac{2^n - 1}{2^n}(1 - \lambda_1) \leqslant p \leqslant \frac{4^n + 2^n - 2}{4^n}(1 - \lambda_1).$$

Using the upper bound as our estimate of p corresponds to assuming that $p_2 = 0$ and leads to overestimating the entanglement infidelity by a factor of at most $(2^n + 2)/2^n$, that is, by a factor of 2 or less. For the purpose of benchmarking logical qubits, however, the estimate will typically be much better since logical Y errors are strongly suppressed for many popular codes due to such errors having higher weight and/or separate syndrome measurement and decoding for physical X and Z errors. In the toric code, for example, twice as many physical errors are required to generate a logical Y error as to generate a logical X or logical Z error.

---

[3]Note that any multiqubit state that is an eigenstate of more than one Pauli operator is an eigenstate of at least one nontrivial real Pauli operator.

B. Controlled-NOT and Pauli gates

Now consider the subgroup of the Clifford group generated by controlled-NOT and the single-qubit Pauli gates. Controlled-NOT gates generate the group GF_2 through their action on Pauli operators containing only X and I elements and separately on Pauli operators containing only Z and I elements. Twirling with respect to this group results in a channel consisting of four blocks

$$\begin{aligned}\tilde{\mathcal{E}}_C(\rho) &= (1-p)\rho + \frac{p_1}{N_1(n)} \sum_{\mu \in \mathbf{B}_1} P_\mu \rho P_\mu \\ &\quad + \frac{p_2}{N_2(n)} \sum_{\mu \in \mathbf{B}_2} P_\mu \rho P_\mu + \frac{p_3}{N_3(n)} \sum_{\mu \in \mathbf{B}_3} P_\mu \rho P_\mu \\ &\quad + \frac{p_4}{N_4(n)} \sum_{\mu \in \mathbf{B}_4} P_\mu \rho P_\mu,\end{aligned}$$

where (i) $\mathbf{B}_1$ consists of the nonidentity Pauli operators containing only Z and I elements. The size of $\mathbf{B}_1$ is

$$N_1(n) = 2^n - 1.$$

(ii) $\mathbf{B}_2$ consists of the nonidentity Pauli operators containing only X and I elements. The size of $\mathbf{B}_2$ is

$$N_2(n) = 2^n - 1.$$

(iii) $\mathbf{B}_3$ consists of the nonidentity Pauli operators containing an even number of Y elements and not belonging to $\mathbf{B}_1$ or $\mathbf{B}_2$. The size of $\mathbf{B}_3$ is

$$N_3(n) = \frac{4^n - 3 \times 2^n}{2} + 1.$$

(iv) $\mathbf{B}_4$ consists of those Pauli operators containing an odd number of Y elements. The size of $\mathbf{B}_4$ is

$$N_4(n) = \frac{4^n - 2^n}{2}.$$

The eigenvalues of $\tilde{\mathcal{E}}_C$ with respect to the Pauli operators of each block are

$$\begin{aligned}\lambda_1 &= 1 - (p_2 + p_3 + p_4)\frac{2^n}{2^n - 1} \\ &= 1 - (p_2 + p_3 + p_4)\left(1 + \frac{1}{2^n}\right) + O(2^{-2n}),\end{aligned}$$

$$\begin{aligned}\lambda_2 &= 1 - (p_1 + p_3 + p_4)\frac{2^n}{2^n - 1} \\ &= 1 - (p_1 + p_3 + p_4)\left(1 + \frac{1}{2^n}\right) + O(2^{-2n}),\end{aligned}$$

$$\begin{aligned}\lambda_3 &= 1 - (p_1 + p_2 + p_4)\frac{2^n}{2^n - 1} - p_3\frac{4^n - 2^{n+2}}{4^n - 3 \times 2^n + 2} \\ &= 1 - p + \frac{p_3 - p_1 - p_2 - p_4}{2^n} + O(2^{-2n}),\end{aligned}$$

$$\begin{aligned}\lambda_4 &= 1 - (p_1 + p_2 + p_3)\frac{2^n}{2^n - 1} - p_4\frac{2^n - 2}{2^n - 1} \\ &= 1 - p + \frac{p_4 - p_1 - p_2 - p_3}{2^n} + O(2^{-2n}).\end{aligned}$$

The ability to prepare and measure eigenstates of X, Y, and Z would enable the reconstruction of all four parameters p_1,

p_2, p_3, and p_4. For logical qubits, at least, preparation and measurement in the X and Z logical bases are often relatively straightforward as most codes of interest are Calderbank-Shor-Steane codes. As discussed in the preceding section, however, preparing and measuring in the logical Y basis is often problematic. The entanglement infidelity averaged over the gates set p can be estimated for two or more qubits by performing benchmarking on eigenstates of Pauli operators in $\mathbf{B}_1$ and eigenstates of Pauli operators in $\mathbf{B}_2$ independently. Given λ_1 and λ_2, the entanglement infidelity can be bounded as follows:

$$\frac{2^n - 1}{2^{n+1}}(2 - \lambda_1 - \lambda_2) \leqslant p \leqslant \frac{2^n - 1}{2^n}(2 - \lambda_1 - \lambda_2).$$

Using the upper bound corresponds to assuming that $p_3, p_4 = 0$ and leads to an overestimate of p by at most a factor of 2 independent of the number of qubits in the benchmarking experiment. For logical qubits the estimate will typically be much better because p_3 and p_4 are likely to be much smaller than p_1 and p_2 for the reasons discussed in the preceding section. Alternatively, for $n > 2$ preparation and measurement of an eigenstate of a Pauli operator in $\mathbf{B}_3$ (e.g., $|{+}00\rangle$) allows λ_3 to be extracted from the decay curve. Given λ_3, the bounds on p are

$$\frac{2^n - 1}{2^n}(1 - \lambda_3) \leqslant p \leqslant \frac{4^n - 3 \times 2^n + 2}{4^n - 2^{n+2}}(1 - \lambda_3),$$

where the upper bound corresponds to taking p_3 to be the only nonzero probability. This is an unrealistic but conservative assumption, causing us to overestimate p by at most a factor of $(2^n - 2)/(2^n - 4)$. For logical qubits the lower bound will typically yield a better estimate of the entanglement infidelity.

IV. CONCLUSION

In this article we have introduced a method for analyzing the behavior of randomized benchmarking as it applies to subgroups of the Clifford group that do not form 2-designs with respect to $\mathrm{SU}(2^n)$. We have additionally applied this method to two subgroups of interest. The first subgroup considered was the real Clifford group, which is generated by controlled-NOT, Hadamard, and Pauli gates. We described a protocol for performing randomized benchmarking on n qubits using only the real Clifford group that estimates the entanglement infidelity of the average error channel to within a factor of $(2^n + 2)/2^n$. The second subgroup considered was that generated by controlled-NOT and Pauli gates. Given the ability to prepare and measure both $|0\rangle$ and $|+\rangle$, we found that the entanglement infidelity can be estimated to within a of factor of either 2 or $(2^n - 2)/(2^n - 4)$ depending on which decay constants are extracted. These results demonstrate that highly accurate approximate randomized benchmarking can be performed using subgroups of the Clifford group which are not unitary 2-designs.

Note added. Recently, we learned of closely related, independent work by Hashagen *et al.* in which they develop a randomized benchmarking protocol for the real Clifford group [23].

[1] K. R. Brown, A. C. Wilson, Y. Colombe, C. Ospelkaus, A. M. Meier, E. Knill, D. Leibfried, and D. J. Wineland, Single-qubit-gate error below 10^{-4} in a trapped ion, Phys. Rev. A **84**, 030303(R) (2011).

[2] J. P. Gaebler, A. M. Meier, T. R. Tan, R. Bowler, Y. Lin, D. Hanneke, J. D. Jost, J. P. Home, E. Knill, D. Leibfried, and D. J. Wineland, Randomized Benchmarking of Multiqubit Gates, Phys. Rev. Lett. **108**, 260503 (2012).

[3] J. M. Chow, J. M. Gambetta, L. Tornberg, J. Koch, L. S. Bishop, A. A. Houck, B. R. Johnson, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Randomized Benchmarking and Process Tomography for Gate Errors in a Solid-State Qubit, Phys. Rev. Lett. **102**, 090502 (2009).

[4] R. Barends *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature (London) **508**, 500 (2014).

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

[6] S. Olmschenk, R. Chicireanu, K. D. Nelson, and J. V. Porto, Randomized benchmarking of atomic qubits in an optical lattice, New J. Phys. **12**, 113007 (2010).

[7] T. Xia, M. Lichtman, K. Maller, A. W. Carr, M. J. Piotrowicz, L. Isenhower, and M. Saffman, Randomized Benchmarking of Single-Qubit Gates in a 2D Array of Neutral-Atom Qubits, Phys. Rev. Lett. **114**, 100503 (2015).

[8] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, K. M. Itoh, D. N. Jamieson, J. C. McCallum, A. S. Dzurak, and A. Morello, Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking, J. Phys.: Condens. Matter **27**, 154205 (2015).

[9] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).

[10] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[11] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, Phys. Rev. Lett. **106**, 180504 (2011).

[12] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).

[13] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).

[14] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Investigating the limits of randomized benchmarking protocols, Phys. Rev. A **89**, 062321 (2014).

[15] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Characterizing universal gate sets via dihedral benchmarking, Phys. Rev. A **92**, 060302 (2015).

[16] A. W. Cross, E. Magesan, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, Scalable randomised benchmarking of non-Clifford gates, Npj Quantum Inf. **2**, 16012 (2016).

[17] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, What Randomized Benchmarking Actually Measures, Phys. Rev. Lett. **119**, 130502 (2017).

[18] J. J. Wallman, Randomized benchmarking with gate-dependent noise, Quantum **2**, 47 (2018).

[19] R. Cleve, D. Leung, L. Liu, and C. Wang, Near-linear constructions of exact unitary 2-designs, Quantum Inf. Comput. **16**, 0721 (2016).

[20] M. A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Phys. Lett. A **303**, 249 (2002).

[21] J. J. Wallman and S. T. Flammia, Randomized benchmarking with confidence, New J. Phys. **16**, 103032 (2014).

[22] J. Helsen, J. J. Wallman, S. T. Flammia, and S. Wehner, Multi-qubit randomized benchmarking using few samples, arXiv:1701.04299.

[23] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, Real randomized benchmarking, arXiv:1801.06121.