Randomized Benchmarking with Confidence

Joel J. Wallman and Steven T. Flammia

The main result of this paper is to prove the first rigorous confidence interval on randomized benchmarking, the *de facto* standard method for characterizing the accuracy of experimental implementations of quantum gates. This theoretical result is important for precisely determining sources of noise in all gate-model-based quantum architectures, and provides an improvement over previous sample complexity estimates (for parameter ranges of interest) by 3 orders of magnitude. The proof techniques involved are also of interest to the QIP community and involve bounding the second moment of ensembles of random Clifford circuits as well as spectral properties of quantum channels. The full version of this extended abstract is available as arXiv:1404.6025.

One of the key obstacles to realizing large-scale quantum computation is the need for error correction and fault tolerance [1], which require the coherent implementation of unitary operations to high precision. Characterizing the accuracy of an experimental implementation of a unitary operation is therefore an important prerequisite for constructing a large-scale quantum computer.

It is possible to completely characterize an experimental implementation of a unitary using full quantum process tomography [2, 3]. However, this approach is impractical for large quantum systems since it involves an exponential amount of resources in the number of qubits and it is sensitive to state preparation and measurement (SPAM) errors, which create a noise floor below which an accurate process estimation becomes impossible [4]. Finally, it does not capture any notion of systematic, time-dependent errors that can arise from applying many unitaries in sequence.

One can avoid the exponential scaling by accepting a partial characterization of an experimental implementation. A partial characterization of, for example, the average error rate and/or the worst-case error rate compared to a perfect implementation of a target unitary is typically enough to determine whether an experimental implementation of a unitary is sufficient for achieving fault-tolerance in a specific scheme for fault-tolerant quantum computation. Such partial characterizations can be obtained efficiently (in the number of quantum systems) using either randomized benchmarking [5–10] or direct fidelity estimation [11, 12].

While direct fidelity estimation gives an unconditional and assumption-free estimate of the average gate fidelity, it is prone to state preparation and measurement (SPAM) errors, which leads to conflation of noise sources. Thus, a key advantage of randomized benchmarking is that it is not sensitive to SPAM errors. Unfortunately, however, current proposals for randomized benchmarking assume that the noise is time-independent, although time-dependence can be partially characterized by a deviation from the expected fidelity decay curve [7, 8]. Furthermore, experimental implementations of randomized benchmarking typically use on the order of 100 random sequences of Clifford gates, which is three orders of magnitude smaller than the number of sequences suggested by the rigorous bounds in Ref. [8] to obtain an accuracy comparable to the claimed experimental accuracies [10, 13]. Numerical investigations of a variety of noise models have shown that between 10–100 random sequences for each length are sufficient to provide a tight estimate of the average gate fidelity [14]. Ideally, one would like to combine the advantages of both randomized benchmarking and direct fidelity estimation to achieve a method that is insensitive to SPAM, requires few measurements, is nearly assumption-free (i.e., does not assume a specific noise model), and comes with rigorous guarantees on the errors involved.

We provide a new analysis of randomized benchmarking which brings it closer in line with this ideal. We first show that the standard protocol can be modified to provide a means of estimating the time-dependent average gate fidelity (which characterizes the average error rate), which also provides an indicator for non-Markovian evolution.

We then provide a *rigorous justification* for taking a small number of random sequences at each length that is on the same order as used in practice by obtaining bounds on the variance due to sampling gate sequences. Numerically, we observe that our bounds (at least for qubits) are saturated, as shown in Fig. 1, and so cannot be improved without further assumptions on the noise (e.g., that the noise is diagonal in the Pauli basis). Therefore any experiments using fewer random sequences than justified by our analysis (unless there is solid evidence that the noise has a specific structure)

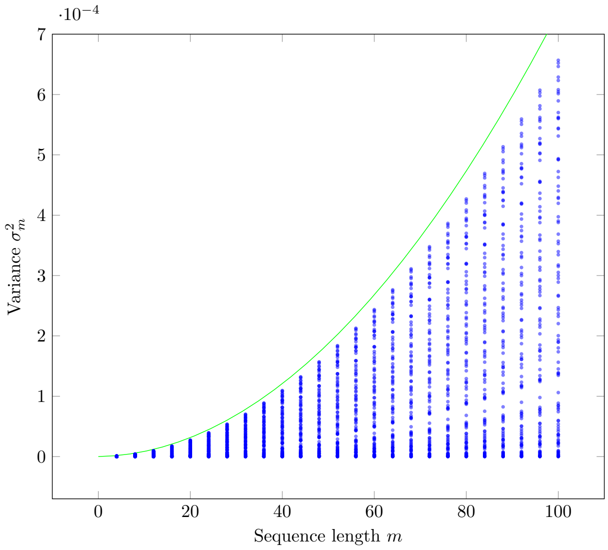

FIG. 1. Plot of benchmarking variance as a function of gate sequence length. Our upper bound on the benchmarking variance (green) is essentially tight for parameter ranges of current experimental interest (blue scatter plots). See main text for details.

will potentially underestimate the error due to sampling random sequences.

In order to give a rigorous statement of results, we will briefly review the randomized benchmarking protocol. The goal of randomized benchmarking is to efficiently but partially characterize the average noise in an experimental implementation of a finite group of operations $\mathcal{G}$ acting on a d-dimensional quantum system. In order to characterize the average noise in an implementation of $\mathcal{G}$ using randomized benchmarking, we require $\mathcal{G}$ to be a unitary 2-design (usually taken to be the Clifford group on n qubits for $d = 2^n$), meaning that sampling over $\mathcal{G}$ reproduces the second moments of the Haar measure [15, 16].

The protocol (see the main text for details) is then to perform several randomly chosen gate sequences of increasing length followed by a single gate that – if gate in the circuit were perfect – inverts the sequence so that the final measurement result is in the same basis as the original state preparation, and hence deterministic. When the gates are noisy, and under reasonable assumptions such as Markovian noise, the variation in the final measurement decreases exponentially with the sequence with a rate given by $F_{m,s}$, the average fidelity at the output given sequence s of length m.

We can regard the probability $F_{m,s}$ as a realization of a random variable F_m with variance σ_m^2. It is this variance that we study in this paper, and which fundamentally affects the accuracy of benchmarking experiments.

Obtaining estimates $\hat{F}_m$ of F_m for multiple m and fitting to the model $\bar{F}_m = A + Bf^m$ will give an estimate of f, the average gate fidelity, provided that the noise does not depend too strongly on the target gate [8], where [17]

$$f = \frac{d\mathcal{F}_{\text{avg}}(\mathcal{E}) - 1}{d - 1} \tag{1}$$

and $\mathcal{F}_{\text{avg}}(\mathcal{E}) = \int \mathrm{d}\psi \mathrm{Tr}\big[\psi\mathcal{E}(\psi)\big]$ is the average gate fidelity of a noise channel $\mathcal{E}$ with respect to the identity channel and $\mathrm{d}\psi$ is the uniform Haar measure over all pure states. The average gate fidelity of $\mathcal{E}$ gives the average probability that preparing a state ψ, applying $\mathcal{E}$ and then measuring $\{\psi, \mathbb{1} - \psi\}$ will give the outcome ψ, averaged over all pure states ψ. That is, $\mathcal{E}$ is the error channel per operation, averaged over all operations in $\mathcal{G}$.

Our first contribution is to show that randomized benchmarking can be used to characterize time-dependent fluctuations in the noise strength. In particular, we demonstrate a data-dependent inequality which, if violated, guarantees the presence of nonMarkovian noise in the system.

Our main contribution is to show that the number of random sequences that need to be averaged is comparable to the number actually used in contemporary experiments (compared to previous best estimates, which require 3 orders of magnitude more random sequences than currently used). In particular, we obtain explicit upper bounds on the variance of the benchmarking distribution of

$$\sigma_m^2 \leq m^2 r^2 + \frac{7mr^2}{4} + 6\delta mr + O(m^2 r^3) + O(\delta m^2 r^2)\,, \tag{2}$$

and

$$\sigma_m^2 \leq 4d(d+1)mr + O(m^2 r^2 d^4)\,. \tag{3}$$

on the variance over random sequences for benchmarking qubits and d-level systems respectively, where for the qubit bound, δ quantifies the deviation from preparations and measurements in a Pauli eigenstate. Applying a variance-dependent version of Hoeffding's inequality [19], sampling $K = -\log(2/\delta)/\log\big[H(\epsilon,\sigma_m^2)\big]$ random sequences is sufficient to obtain an absolute precision of ϵ in the estimate of $\bar{F}_m$ with probability $1-\delta$, where $H(\epsilon,v) = \left(\frac{1}{1-\epsilon}\right)^{\frac{1-\epsilon}{v+1}}\left(\frac{v}{v+\epsilon}\right)^{\frac{v+\epsilon}{v+1}}$. To illustrate the resources required for a rigorously guaranteed precision using our bounds, consider the following parameters for a single-qubit benchmarking experiment: $m = 100, r = 10^{-4}, \epsilon = 1\%, \delta = 1\%$. For these parameters, our upper bound of $\sigma_m^2 = m^2r^2 + \frac{7}{4}mr^2$ (ignoring the higher-order terms) shows that $K = 145$ random sequences suffice. This is an improvement by *orders of magnitude* over the previous best rigorously justifiable upper bound of 10^5 using the variance-independent Hoeffding inequality [8].

Importantly, however, the quadratic scaling with m in the regime $mr \ll 1$ is necessary for general noise source (see Fig. 1). Therefore, a corollary of our result is that longer sequence lengths should be averaged over more random sequences in this regime, since having larger variance for larger m would generally cause less weight to be assigned to larger m when fitting to extract the average gate fidelity.

Furthermore, we prove that if the noise is unitary (or nonunital), then the variance due to sampling random sequences converges to a constant. If such noise sources (including nonunital noise and any unitary noise, such as over- and under-rotations) are believed to be present, substantially more random sequences need to be sampled. As such, our analysis shows that randomized benchmarking is most reliable in the regime $mr \ll 1$, although, since the next lowest order terms in our bound are $\delta m^2 r^2$ and $m^2 r^3$, the lowest order bounds on the variance should be approximately valid for $mr \approx 0.1$.

In conclusion, our results show rigorously that randomized benchmarking with arbitrary Markovian noise is generically *almost* as accurate as has previously been estimated in experiments [9, 10, 13] and numerics [14]. However, we find that the number of random sequences should scale with the sequence so that the variance is independent of m. We also find that if near-unitary noise (such as under- and over-rotations) are a predominant noise source, then randomized benchmarking should be conducted in the regime $mr \ll 1$, since the variance due to sampling random sequences with such noise sources will remain large as m increases.

---

[1] Daniel Gottesman, "An Introduction to Quantum Error Correction and Fault-Tolerant Quantum Computation," (2009), arXiv:0904.2557v1.

[2] Isaac L. Chuang and Michael A. Nielsen, "Prescription for experimental determination of the dynamics of a quantum black box," J. Mod. Opt. **44**, 2455 (1997).

[3] J. Poyatos, J. Cirac, and P. Zoller, "Complete Characterization of a Quantum Process: The Two-Bit Quantum Gate," Phys. Rev. Lett. **78**, 390 (1997).

[4] 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," Phys. Rev. A **87**, 062119 (2013).

[5] Joseph Emerson, Robert Alicki, and Karol Życzkowski, "Scalable noise estimation with random unitary operators," J. Opt. B Quantum Semiclassical Opt. **7**, S347 (2005).

[6] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. 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).

[7] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, "Scalable and Robust Randomized Benchmarking of Quantum Processes," Phys. Rev. Lett. **106**, 180504 (2011).

[8] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, "Characterizing quantum gates via randomized benchmarking," Phys. Rev. A **85**, 042311 (2012).

[9] 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).

[10] 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, Mark B. Ketchen, and M. Steffen, "Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking," Phys. Rev. Lett. **109**, 080505 (2012).

[11] Marcus P. da Silva, Olivier Landon-Cardinal, and David Poulin, "Practical Characterization of Quantum Devices without Tomography," Phys. Rev. Lett. **107**, 210404 (2011).

[12] Steven T. Flammia and Yi-Kai Liu, "Direct Fidelity Estimation from Few Pauli Measurements," Phys. Rev. Lett. **106**, 230501 (2011).

[13] 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 (2011).

[14] Jeffrey M Epstein, Andrew W Cross, Easwar Magesan, and Jay M Gambetta, "Investigating the limits of randomized benchmarking protocols," (2014), arXiv:1308.2928v1.

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

[16] David Gross, K. Audenaert, and Jens Eisert, "Evenly distributed unitaries: On the structure of unitary designs," J. Math. Phys. **48**, 052104 (2007).

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

[18] B. R. Johnson, C. A. Ryan, M. P. da Silva, S. Kimmel, and T. Ohki, in preparation (2014).

[19] Wassily Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," J. Am. Stat. Assoc. **58**, 13 (1963).