Statistical analysis of randomized benchmarking

Robin Harper, Ian Hincks, Chris Ferrie, Steven T. Flammia, and Joel J. Wallman

Phys. Rev. A **99**, 052350 — Published 29 May 2019

DOI: 10.1103/PhysRevA.99.052350

Statistical analysis of randomized benchmarking

Robin Harper,[1] Ian Hincks,[2,3] Chris Ferrie,[4] Steven T. Flammia,[1,5] and Joel J. Wallman[2]

[1] *Centre for Engineered Quantum Systems, School of Physics, The University of Sydney, Sydney, Australia*
[2] *Institute for Quantum Computing and Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*
[3] *Quantum Benchmark Inc., Kitchener, ON N2H4C3, Canada*
[4] *Centre for Quantum Software and Information, University of Technology Sydney, Australia*
[5] *Yale Quantum Institute, Yale University, New Haven, CT 06520, USA*

(Dated: May 1, 2019)

Randomized benchmarking and variants thereof, which we collectively call RB+, are widely used to characterize the performance of quantum computers because they are simple, scalable, and robust to state-preparation and measurement errors. However, experimental implementations of RB+ allocate resources suboptimally and make ad-hoc assumptions that undermine the reliability of the data analysis. In this paper, we propose a simple modification of RB+ which rigorously eliminates a nuisance parameter and simplifies the experimental design. We then show that, with this modification and specific experimental choices, RB+ efficiently provides estimates of error rates with multiplicative precision. Finally, we provide a simplified rigorous method for obtaining credible regions for parameters of interest and a heuristic approximation for these intervals that performs well in currently relevant regimes.

I. INTRODUCTION

Characterizing large scale quantum devices is a prerequisite to optimizing their performance and being able to reliably perform useful information processing tasks. Full characterization is manifestly not scalable for general errors, so that scalable methods can only partially characterize the noise. Currently, the only fully scalable protocols that partially characterize quantum devices are randomized benchmarking [1–4] and variants thereof (RB+) [5–11]. This family of protocols can provide a wide variety of information about noise parameters, including the average error rate [2–4, 12], error rates for specific gates [5, 6, 10, 11], leakage rates [13], loss rates [8, 12], and the amount of residual unitary (calibration) errors [7, 9, 14].

RB+ provides estimates of noise parameters by applying long sequences of random gates to amplify errors in the implementation of gates and estimate them independently from state preparation and measurement errors (SPAM). Typically, descriptions of RB+ state that experiments should be repeated to obtain a desired precision without necessarily specifying (or recommending) any of the following: (1) estimators for finite data; (2) how many repetitions should be performed; or (3) how finite and heteroscedastic data should be fit to a specified model. This last point is important because RB+ data are generally heteroscedastic, meaning that the variance across the data is non-uniform, since the variance over random sequences increases with the sequence length [15]. Abstaining from specifics on these points was perhaps warranted by the fact that particular choices are difficult to derive or justify as being optimal, or nearly optimal. Obtaining a fully general and optimal specification is confounded by the unknown distribution of errors over the random sequence of gates [15, 16]. However, in one experimental regime, Bayesian techniques can be applied to obtain rigorous credible intervals for the model parameters as well as efficient allocation of experimental measurements [17, 18]. We discuss and utilize this work by Granade et al. in section IV.

In this paper, we present a minor modification of RB+ that improves the efficiency of the method by eliminating a nuisance model parameter, yet it adds no experimental overhead. Similar methods have been presented previously in the literature for the case of single-qubit RB [19, 20]. For the two remaining model parameters, we then provide estimators which do not have to be weighted to correct for heteroscedasticity because they can be estimated from two independent sequence lengths. Without our modification, at least 3 sequence lengths are required, which in turn require a weighted fit where the correct weights are not generally inferable from the data. We then study the distribution of the parameter estimators and show how to obtain simple and rigorous credible intervals in the regime studied in Ref. [17], that is, when each random sequence is repeated once. Finally, we also provide a simple proof that certain experimental design choices enable RB+ to efficiently provide estimates of error rates that have multiplicative precision. By showing that the estimates of such error rates have multiplicative precision, we confirm that RB+ will continue to allow efficient estimation of the model parameters as gate fidelity rates improve through the simple expedient of increased sequence lengths.

In particular we provide guidance as to the two remaining matters typically left vague in previous papers on RB+, namely:

1. *Gate lengths*: For the purpose of fidelity estimation we find that experimental effort should be concentrated on

two sequence lengths ($m = 4$) and $m = 1/[2(1-p)]$ (section III). Where there is no prior estimate of p, then sequences doubling in length (see Algorithm 1) can be used. From the results of the experiment the correct gate length sequence with which to perform the fit can be selected (i.e. the one which has approximately 1/3 of the survival probability of the first gate). We note that whilst the fit should concentrate on just these two sequences it would be prudent to measure survival probability at other sequence lengths, to ensure an exponential decay for the purposes of model validation (for example, to ensure the system is not afflicted by model breaking non-Markovian noise).

2. *Number of sequences*: We show that with a probability of $(1-\delta)$, one can obtain a multiplicative error ϵ on the infidelity estimation, using $O\left(\frac{1}{\epsilon^2}\log(\frac{1}{r})\log[\frac{1}{\delta}\log(\frac{1}{r})]\right)$ measurements. For instance assuming an infidelity of 5×10^{-3} then then a 2% relative error in the estimated fidelity can be achieved with a 95% confidence level by sampling on order of 100 sequences with 1024 shots per sequence. Further, more detailed, examples are given in Figure 1 and the proof is in Section V. Typically to avoid the estimator being biased a very small number of sequences should not be used (see section III).

It is important to note that all of our error estimates and sequence number results rely on an appropriate sequence length being chosen. If sequences lengths are too short then the number of sequences needed to obtain accurate estimates will increase dramatically, if sequences are too long then inaccurate estimates will be obtained. Algorithm 1 provides a systematic way to ensure the correct sequence lengths are chosen.

In what follows, we use the notation that $\hat{x}$ is an estimator of a quantity $\bar{x}$, where the bar denotes that either an expected value or a sample average has been taken over realizations of a random variable x.

II. RB+ PROTOCOL

We begin by providing a general framework that describes all existing RB+ protocols except the leakage protocol of Ref. [21] and the unitarity protocol of Ref. [7]. We will also present a modified version of the unitarity protocol. We exclude the leakage protocol of Ref. [21] because it has a more complicated fit model that is not robust to SPAM errors.

RB+ protocols are of the following form.

1. Choose a positive integer m.
2. Choose a random sequence of gates s from a set $\mathbb{S}_m$, typically of Clifford gates. Note that these gates are often chosen to leave a state invariant. However, as discussed below, uniformly choosing s to either leave the state invariant or map it to an orthogonal state eliminates a nuisance model parameter.
3. Obtain an estimate $\hat{q}(m,s)$ of the expectation value $q(m,s)$ of an observable E after preparing a state ρ and applying the gates in s. Typically ρ should be close to an ideal computational basis state and E should be close to a projector onto a pure state in the computational basis.
4. Repeat steps 2–3 k_m times to obtain an estimate $\hat{q}(m)$ of $\bar{q}(m) = |\mathbb{S}_m|^{-1}\sum_{s\in\mathbb{S}_m} q(m,s)$.
5. Repeat steps 1–4 and fit to the model

$$\bar{q}(m) = Ap^m + B \tag{1}$$

where p is related to some parameter of interest (e.g., the average gate fidelity to the identity) and A and B are SPAM-dependent constants.

We assume throughout that $A \gg 0$, which holds in current regions of interest, as otherwise it is unclear how to efficiently gather useful statistics.

A. Unitarity

We now introduce a slight variant of the RB+ protocol to handle the special case of the unitarity protocol from Ref. [7]. The variant enables an independent estimate of the unitarity (which quantifies how coherent the errors are) and the leakage rate [8, 13]. Note that the following protocol is not strictly scalable as it involves sampling every

Pauli matrix and also assumes that there is no (or minimal) state-dependent loss. The scalability could be improved by, for example, performing importance sampling of the Pauli matrices conditioned on the sequence [22]; however, we leave this as an open problem.

1. Choose a positive integer m.
2. Choose a random sequence of m n-qubit Clifford gates s.
3. For each n-qubit Pauli matrix P, obtain an estimate $\hat{q}(m, s|P)$ of the expectation value of the observable P after preparing a fixed state ρ and applying the gates in s.
4. Repeat steps 2–3 k_m times. For each n-qubit Pauli matrix $P \neq I$, set

$$\begin{aligned} \hat{a}(m|P) &= \sum_s \hat{q}(m, s|P)/k_m \\ \hat{a}(m) &= \frac{1}{4^n - 1} \sum_P \hat{a}(m|P) \\ \hat{b}(m) &= \frac{1}{k_m} \sum_{P,s} \hat{q}(m, s|P)^2 - \hat{a}(m|P)^2. \end{aligned} \tag{2}$$

5. Repeat steps 1–4 and fit to the models

$$\begin{aligned} \bar{a}(m) &= Al^m \\ \bar{b}(m) &= A'u^m \end{aligned} \tag{3}$$

where l and u are the leakage rate [13] and the unitarity [7] respectively.

Unlike other protocols, the combined unitarity/loss protocol does not require any truncation of $\bar{q}(m)$ to avoid negative values. We also note that recently an alternative protocol has been proposed which proposes a method for efficient unitarity benchmarking in the regime of few-qubit Clifford gates [23].

B. Eliminating the offset

While superficially benign, since it is a free fitting parameter, the variable offset B in eq. (1) can severely increase the marginal uncertainty in p, the parameter of interest [19]. We now present a method of rigorously and exactly eliminating this constant offset without having to estimate its value.

First note that for gate independent noise Λ, the constant

$$B := \mathrm{Tr}\big[E\Lambda(\mathbb{1}/2^n)\big] \tag{4}$$

and the decay parameter p [4, 12] do not change if we compile any gate into the sequence. (In fact, strictly this holds even for non-trace-preserving noise where B is multiplied by a second exponential). In particular, let X be any gate that maps the input state to an orthogonal state, as the single-qubit Pauli X operator does for states in the computational basis. Let $\mathbb{S}_{m,b}$ be the set of sequences obtained from $\mathbb{S}_m$ by compiling X^b into the sequence and let

$$\hat{q}(m|b) = \frac{1}{|\mathbb{S}_{m,b}|} \sum_{s \in \mathbb{S}_{m,b}} q(m, s). \tag{5}$$

Then we have

$$\bar{q}(m) = \bar{q}(m|0) - \bar{q}(m|1) = Ap^m \tag{6}$$

where now $A \in [0, 1]$. A similar idea was suggested for single qubits in [19, 20]. One disadvantage of this approach is that the remaining A coefficient in eq. (1) may be small for some values of b, especially for multiple qubits, thus reducing the signal from some experiments.

Alternatively, consider an n-qubit POVM $\{E_1, \ldots, E_k\}$ and suppose that a set of gates $\{X_1, \ldots, X_k\}$ are such that $E_j \approx X_j E_1 X_j^\dagger$. Then by compiling X_j into the sequence uniformly at random and recording the probability of observing the corresponding E_j and averaging over j, the average value of B becomes

$$B = \frac{1}{k} \sum_{j=1}^{k} \mathrm{Tr}\big[E_j \Lambda(\tfrac{\mathbb{1}}{2^n})\big] = \frac{1}{k} \mathrm{Tr}\big[\mathbb{1}\Lambda(\tfrac{\mathbb{1}}{2^n})\big] = \frac{1}{k}. \tag{7}$$

III. ESTIMATING THE DECAY RATE

With a known value of B, eqs. (1) and (3) have two unknown parameters and so we need at least two values of m, denoted $m_1 < m_2$, to estimate either (or both) parameters.

From eq. (1),

$$\begin{aligned} A &= [\bar{q}(m_1) - B]^{m_2/\delta m} \, [\bar{q}(m_2) - B]^{-m_1/\delta m}, \\ p &= [\bar{q}(m_1) - B]^{-1/\delta m} \, [\bar{q}(m_2) - B]^{1/\delta m}. \end{aligned} \tag{8}$$

Each of these terms is of the form $x_1^{\alpha_1} x_2^{\alpha_2}$ where $x_j^\alpha = [\bar{q}(m_j) - B]^\alpha$. A natural approach would be to estimate x_j^α by $[\hat{q}(m_j) - B]^\alpha$, where $\hat{q}(m_j)$ is an unbiased estimator for $\bar{q}(m_j)$ (e.g., the sample mean). However, this approach has two issues. First, the estimator is complex or undefined if $\hat{q}(m_j) \leq B$. To address this issue, we can truncate $\bar{q}(m_j)$ to $B + \delta$ for some fixed $0 < \delta \ll 1$, which will occur with negligible probability provided enough sequences are taken and the $\hat{q}(m_i)$ are sufficiently far from 0.

A second issue is that if $\hat{q}(m_j)$ is an unbiased estimator for $\bar{q}(m_j)$, then $[\hat{q}(m_j) - B]^\alpha$ is a biased estimator of x_j^α for any $\alpha \neq 1$. To address this second issue, we can estimate and then subtract the bias if necessary. Repeating steps 2–4 for a fixed m and using final gates compiled in to remove the offset B yields an estimate $\hat{q}(m_j) - B = x_j(1 + \epsilon_j)$ for some random variable ϵ_j with zero mean. We then have

$$\begin{aligned} \mathbb{E}\,[\hat{q}(m_j) - B]^\alpha &= x_j^\alpha \mathbb{E}(1 + \epsilon_j)^\alpha \\ &= x_j^\alpha \left[1 + \tfrac{1}{2}\alpha(\alpha - 1)\mathbb{E}\epsilon_j^2 + O(\mathbb{E}\alpha\epsilon^3)\right] \\ &\approx x_j^\alpha + \tfrac{1}{2}\alpha(\alpha - 1) x_j^{\alpha-2} \mathbb{V}\hat{q}(m_j). \end{aligned} \tag{9}$$

Therefore $[\hat{q}(m_j) - B]^\alpha$ is biased but consistent and so the bias can be neglected when sufficiently many sequences are sampled at each m_j. Moreover, the bias is modulated by α, which as we prove below in section V, is $O(r)$ in the optimal regime for p. For small numbers of sampled sequences, the bias term can be subtracted using sample estimates of x_j^α and $\mathbb{V}\hat{q}(m_j)$ on the right-hand-side of eq. (9). As our numerics will show in section IV, the bias is negligible for intermediate numbers of measurements but is noticeable for very small numbers of sequences.

To determine approximately optimal values of m_1 and m_2 assuming that the bias in eq. (9) is negligible, note that

$$\mathbb{V}\,[\hat{q}(m_j) - B]^\alpha = x_j^{2\alpha-2} \alpha^2 \mathbb{V}[\hat{q}(m_j)] + O(\mathbb{E}\alpha^3\epsilon^3). \tag{10}$$

Now we note that by Chebyshev's inequality,

$$\Pr\left(|\hat{p} - p| > k\sqrt{\mathbb{V}(\hat{p})}\right) \leq k^{-2}, \tag{11}$$

so that with probability 8/9 (for example) and using the trivial bound $p \leq 1$, we have

$$|\hat{p} - p| \leq \frac{3}{\delta m}\left(\sum_j \mathbb{V}[\hat{q}(m_j)]\right)^{1/2} + O\big(\mathbb{E}(\epsilon_j^{3/2})/\delta m^{3/2}\big). \tag{12}$$

The only unknowns in eq. (12) are the variances at the two sequence lengths. Choosing $\delta m \approx 1/(1-p)$ therefore gives a multiplicative precision estimate of the error rate $1 - p$, as claimed.

We would like to make the error term as small as possible. To achieve a large δm, and hence a small error, we want m_1 as small as possible and m_2 as large as possible. However, eq. (1) is only typically accurate for $m \geq 4$ [12], so we henceforth set $m_1 = 4$. Furthermore, the number of sequences required to make the truncation probability negligible increases with m_2, and eq. (10) is inversely proportional to x_j for $\alpha \leq 1$, so m_2 can only be increased to some fixed value. Thus, when sampling to some fixed constant accuracy we must not choose m_2 to be too large.

Ref. [17] recommended $m_2 = \lceil 1/(1-p) \rceil$ as the optimal choice of m_2, however, this was for $B = 0$ (i.e., the infinite-dimensional limit). Numerically, we observe that $m_2 = \lceil 1/[2(1-p)] \rceil$ results in a more precise estimate.

Consequently, provided gate lengths (m) can be increased as specified above, the number of sequences required to determine $1 - p$ to within a specified factor remain approximately independent of p. This demonstrates that RB+ protocols scale favorably with the error rate.

This derivation is not quite rigorous only for a very trivial reason, namely because our use of Taylor's theorem requires that we control the smoothness of the functions being expanded over some region, and this region should also be appropriately defined. These expressions are nonetheless useful for practical analysis of RB+. By contrast, the derivation in section V is completely rigorous, but the proof is not meant to provide anything more than coarse guidance about how to choose parameter settings in practical situations.

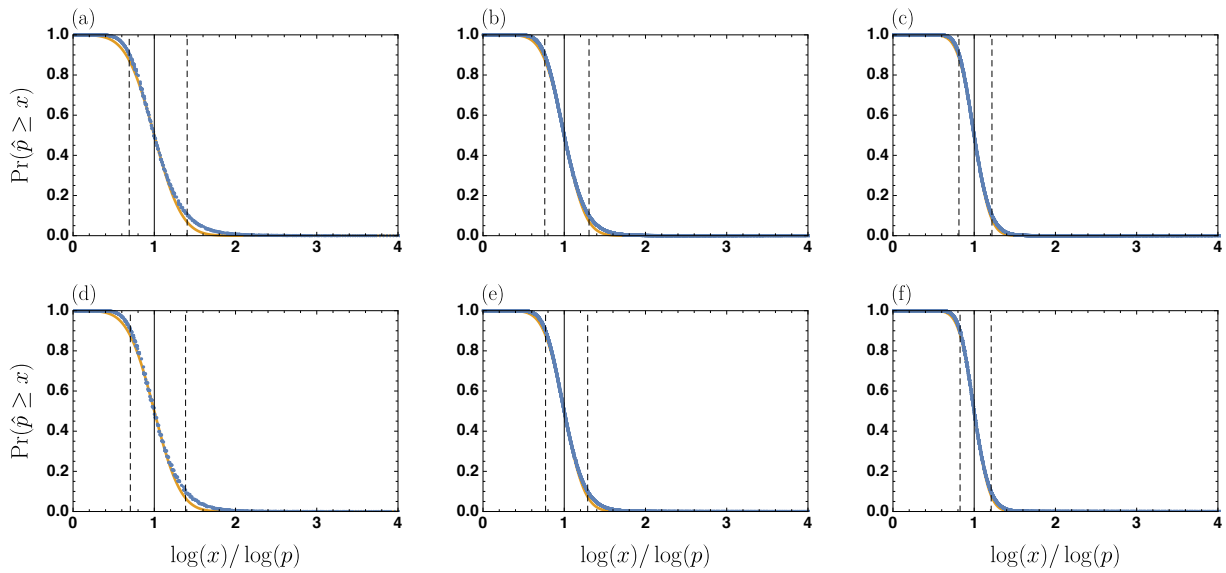

FIG. 1. Numerical demonstration that error rates can be accurately estimated with multiplicative precision using minimal resources. We plot the cumulative density function (CDF) for the estimator $\hat{p}$ from binomial statistics (dots upper/blue) and eq. (14) (solid line lower/orange) for different values of A, p and k. The values used for each of the subfigures are as follows: for (a)—(c), $p = 0.99$ with (a) $A = 0.49\,, k_1 = k_2 = 100$, (b) $A = 0.39\,, k_1 = k_2 = 300$ and (c) $A = 0.49\,, k_1 = k_2 = 100$. For subfigures (d)–(f), $p = 0.9999$ with (d) $A = 0.49\,, k_1 = k_2 = 100$, (e)$A = 0.39\,, k_1 = k_2 = 300$ and (f) $A = 0.49\,, k_1 = k_2 = 300$. Note that the shape is essentially independent of p but that smaller values of A (that is, larger state-preparation and measurement errors) result in heavier tails. Vertical dashed lines are located at the 10% and 90% quantiles of the CDF of the binomial distribution.

IV. ACCELERATED RB

We now analyze accelerated randomized benchmarking (ARB) [17] using the modified protocol discussed in section II B. In ARB, each individual estimate $\hat{q}(m, s)$ is in $\{0, 1\}$, that is, each random sequence is measured once. Therefore, $\hat{q}(m) \sim \mathcal{B}(k_m, \bar{q}(m))/k_m$, where $\mathcal{B}(n, p)$ denotes the binomial distribution with n trials and probability p.

For sufficiently many samples, log ratios of binomial variables are approximately normally distributed [24], so that

$$\begin{aligned} \log \tfrac{\hat{q}(m_2)}{\hat{q}(m_1)} &\sim \mathcal{N}\left(\log(\tfrac{\bar{q}(m_2)}{\bar{q}(m_1)}), \sigma^2\right) \\ \sigma^2 &= \sum_j \frac{\bar{q}(m_j)\,(1 - \bar{q}(m_j))}{k_j(\bar{q}(m_j) - B)^2}. \end{aligned} \tag{13}$$

Therefore

$$\log \hat{p} = \frac{1}{\delta m} \log \tfrac{\hat{q}(m_2)}{\hat{q}(m_1)} \sim \mathcal{N}(\log p, \sigma^2/\delta m^2). \tag{14}$$

In fig. 1, we illustrate that the normal approximation is sufficiently accurate by comparing the exact cumulative density function for the estimator $\hat{p}$ from binomial statistics with the normal approximation of eq. (14) for multiple values of A, B, and p. Note in particular that the shape of the cumulative density function for the estimator $\hat{p}$ is essentially independent of p, but has heavier tails for smaller values of A.

Under the log-normal approximation for $\hat{p}$, the value of m_2 that minimizes the variance of the estimate is given by

$$\operatorname*{argmin}_{m_2} \left(\log\left[p^{-2m_1}\bar{q}(m_1)\,(1 - \bar{q}(m_1)) + p^{-2m_2}\bar{q}(m_2)\,(1 - \bar{q}(m_2))\right] - 2\log(m_2 - m_1)\right). \tag{15}$$

While this minimizes the variance of $\log \hat{p}$ rather than $\hat{p}$, since p near 1, we have $\mathbb{V}(\log \hat{p}) \approx \mathbb{V}(\hat{p})$. This minimization can be performed numerically using an initial value of $-1/\log p$. In fig. 2, we plot the variance and optimal m_2 values in several relevant parameter regimes. The optimal value of m_2 depends on the true values of A, B, and p. In fig. 2(a-b) we see that when choosing a future experiment based on present knowledge with multiplicative uncertainty in p, it is best to err on the side of m_2 that is short with respect to the optimal value.

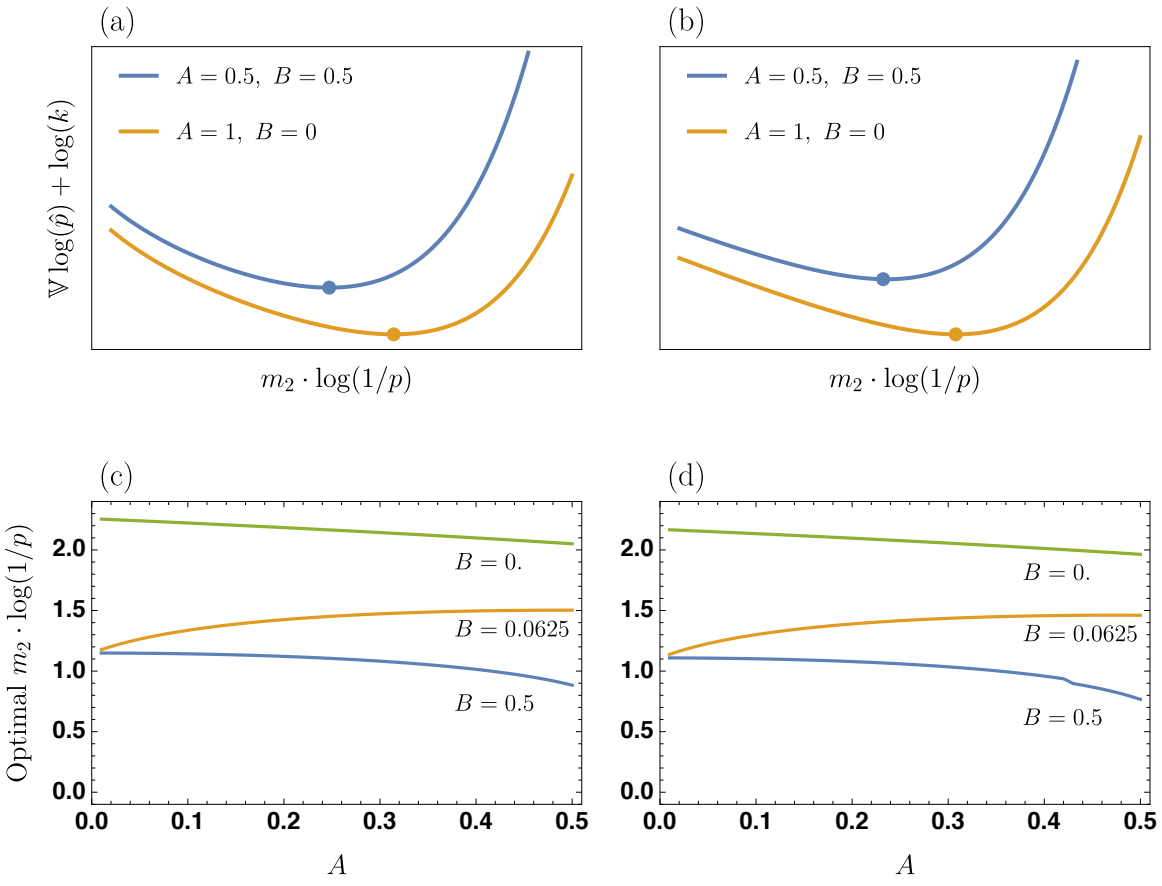

FIG. 2. Illustration that the variance only depends weakly on the choice of m_2 near the optimal value. Subfigures (a) and (b) show the variance of the estimated error rate in eq. (14) plotted for several combinations of A, B, and p, with points placed at the minimum of each curve. In subfigures (a) and (c) $p = 0.99$, and in subfigures (b) and (d) $p = 0.9999$. Note that the x-axis is scaled logarithmically and such that the nominal value $m_2 = 1/\log(1/p)$ appears at the value $x = 1$. Subfigures (c) and (d) show the value of m_2 which minimizes the variance, eq. (15), is plotted as a function of A for several values of B and p. We see that the dependence on p is essentially negligible for practical purposes, and that the dependence on A is fairly weak.

V. RIGOROUS PROOF OF MULTIPLICATIVE PRECISION FOR THE RATIO ESTIMATOR

In this section we give a rigorous proof that RB converges to an estimate with multiplicative precision using the ratio estimator in eq. (8). Here the focus is not on obtaining tight answers, but on having a simple and clear statement of the scaling of the precision that can be achieved assuming the decay model in eq. (6), and achieving a given sample complexity. We therefore largely neglect to track estimation errors closely, focusing instead on the simplest proof possible and a big-O estimate of the resources required.

Let $r = 1 - p$ and fix some small $1/16 > \epsilon_0 > 0$. We are most interested in the regime where r is small, or equivalently p is close to 1. We assume that we can estimate the quantities $q_i = Ap^{m_i}$ with an unbiased estimator $\hat{q}_i = q_i + Ap\epsilon_i$ where the estimation error ϵ_i is a random variable with zero mean. Note that this is multiplicative precision for the case $m_i = 1$, but for larger values of m_i we have just rescaled an additive precision by Ap for algebraic convenience. If the estimator is the sample mean of t Bernoulli random variables with mean q_i, then $\epsilon_i = O(1/\sqrt{t})$ with high probability.

Under these conditions, we can estimate $\hat{p}$ using the following algorithm.

Algorithm 1: Ratio estimator for exponential regression.

1. Set $i := 1$ and $m_1 := 1$.
2. Estimate $\hat{q}_1 := q_1(1 + \epsilon_1)$ using t samples.
3. While $\hat{q}_i > \frac{1}{3}\hat{q}_1$, Do

- Set $i := i + 1$,
- Set $m_i := 2^i + 1$,
- Estimate $\hat{q}_i := q_i + Ap\epsilon_i$ using t samples.

4. Set $\ell := i$ and $m = 2^\ell$.
5. Return $\hat{p} := \left(\frac{\hat{q}_\ell}{\hat{q}_1}\right)^{1/m}$ and $\hat{r} := 1 - \hat{p}$.

We now rigorously prove that the above algorithm returns an estimator with multiplicative precision of $\hat{r}$.

Theorem 1. *For any sufficiently small $\epsilon_0 > 0$, the algorithm above returns estimates $\hat{r}$ such that $|\hat{r} - r| \leq O(\epsilon r)$ with probability $1 - \delta$ using*

$$M = O\left(\frac{1}{\epsilon^2} \log\left(\tfrac{1}{r}\right) \log\left[\tfrac{1}{\delta} \log\left(\tfrac{1}{r}\right)\right]\right) \tag{16}$$

measurements.

The proof relies on a few simple lemmas, which we now state and prove.

Lemma 2. *Given a set of ℓ independent estimates $\hat{q}_i$ obtained from sampling each t times as described above, the probability that $|\hat{q}_i - q_i| \geq Ap\epsilon$ for any $i > 1$ or $|\hat{q}_1 - q_1| \geq q_1\epsilon$ is at most δ if we choose $t = O\left(\frac{1}{\epsilon^2} \log \frac{\ell}{\delta}\right)$, where the implied constant depends on Ap.*

Proof. The proof is an elementary application of the Chernoff bound and the union bound. We omit the details. □

Thus, we can assume that each of the random estimates ϵ_i satisfies $|\epsilon_i| \leq \epsilon$ in the algorithm, and we will fail with probability at most δ. Next, we will see that the algorithm converges in a small number of steps ℓ, and with m taking a value that scales like $1/r$.

Lemma 3. *With probability at least $1 - \delta$, the above algorithm converges with $\ell = \Theta\left(\log \frac{1}{r}\right)$ using $O\left(\frac{\ell}{\epsilon^2} \log \frac{\ell}{\delta}\right)$ total samples, and with m such that*

$$\frac{(1 - 4\epsilon)^2}{9} < p^m \leq \frac{1 + 4\epsilon}{3}. \tag{17}$$

Proof. The algorithm exits the while loop when

$$\hat{q}_\ell = Ap(p^{2^\ell} + \epsilon_\ell) \leq \tfrac{1}{3}\hat{q}_1 = \frac{Ap}{3}\left(1 + \epsilon_1\right). \tag{18}$$

As the algorithm did not exit for $i = \ell - 1$,

$$\hat{q}_{\ell-1} = Ap(p^{2^{\ell-1}} + \epsilon_{\ell-1}) > \tfrac{1}{3}\hat{q}_1 = \frac{Ap}{3}\left(1 + \epsilon_1\right). \tag{19}$$

Squaring the latter inequality and then rearranging both to be in terms of p^m with $m = 2^\ell$, we have

$$\frac{(1 - 3\epsilon_{\ell-1} + \epsilon_1)^2}{9} < p^m \leq \frac{1 - 3\epsilon_\ell + \epsilon_1}{3}. \tag{20}$$

Supposing that for some fixed $\epsilon > 0$, $|\epsilon_i| \leq \epsilon$ for all i with with probability $1 - \delta$, the claim about p^m follows by taking the worst-case choices of the ϵ_i. Taking logarithms of this and using $p = 1 - r$, we find that for any sufficiently small ϵ we have $\ell = \Theta\left[-\log\left(-\log(1 - r)\right)\right]$. As long as r is bounded away from 1 then $-\log(1 - r) = \Theta(r)$, and this is equivalent to $\ell = \Theta\left(\log \frac{1}{r}\right)$.

If we sample as per lemma 2, then $|\epsilon_i| \leq \epsilon$ for all i with probability $1 - \delta$. Therefore the claim about the total number of samples follows immediately from lemma 2. □

Now we are ready to prove the main theorem.

Proof of Theorem. From the above lemmas, we know that with probability $1-\delta$ the ratio estimator converges with $m = O\big(\log \frac{1}{r}\big)$ and errors bounded by ϵ in the numerator and denominator. We have the bounds $\hat{p}_- \leq \hat{p} \leq \hat{p}_+$, where

$$\hat{p}_\pm := \left(\frac{q_\ell \pm Ap\epsilon}{q_1 \mp Ap\epsilon}\right)^{1/m} = p\left(\frac{1 \pm \epsilon/p^m}{1 \mp \epsilon}\right)^{1/m}. \tag{21}$$

From the inequality in eq. (17), we have

$$\hat{p}_+ \leq p\left(1 + \frac{2\epsilon\left(5 - 4\epsilon + 8\epsilon^2\right)}{(1-4\epsilon)^2(1-\epsilon)}\right)^{1/m} \quad \text{and} \quad \hat{p}_- > p\left(1 - \frac{2\epsilon\left(5 - 4\epsilon + 8\epsilon^2\right)}{(1-4\epsilon)^2(1+\epsilon)}\right)^{1/m}. \tag{22}$$

Now we choose any $\epsilon_0 < 1/16$ so that the ϵ dependent terms above are $O(\epsilon)$ and the term for $\hat{p}_-$ remains less than 1. Explicitly evaluating the ϵ dependent terms, we have the bounds

$$\hat{p}_+ < p\big(1 + O(\epsilon)\big)^{1/m} \quad \text{and} \quad \hat{p}_- > p\big(1 - O(\epsilon)\big)^{1/m}, \tag{23}$$

where the implied constant decreases with ϵ_0. Now Taylor expanding in $1/m$ and using the result from lemma 3 that $\ell = \log_2 m = \Theta\big(\log \frac{1}{r}\big)$, we find that

$$\hat{p}_+ < p\big(1 + O(\epsilon q)\big) \quad \text{and} \quad \hat{p}_- > p\big(1 - O(\epsilon r)\big). \tag{24}$$

Adopting the bounds $\hat{r}_\pm = 1 - \hat{p}_\mp$ gives the analogous result for $\hat{r}$. This establishes that the estimator has multiplicative precision,

$$|\hat{p} - p| = |\hat{r} - r| \leq O(\epsilon r)\,. \tag{25}$$

The statement about complexity follows directly from the lemmas, and the theorem is proven. □

VI. CONCLUSION

We have provided a modification to RB+ and concrete recommendations for how to obtain precise estimates of error rates in practical regimes. We have rigorously shown that the precision is multiplicative, and our derivations and numerics demonstrate the utility of the heuristics that we use. Our recommendations are based upon the assumption that the model in eq. (1) is correct. For standard randomized benchmarking, there are only two factors that can cause a deviation from eq. (1) for sequence lengths $m \geq 4$, namely, noise that is time-dependent or non-Markovian [12, 15, 25, 26]. Both types of noise are ubiquitous in experiments and neither can be detected using only two sequence lengths. A standard approach is to take data from more sequence lengths, perform a joint fit and then use the goodness-of-fit as an indicator for non-Markovian noise or drift. However, fitting more sequence lengths is nontrivial as the data are heteroscedastic. Furthermore, adding more sequence lengths does not significantly increase the quality of the error estimates when eq. (1) is correct, and so performing a joint fit provides little extra information and introduces correlations between model estimation and model validation. We instead recommend fitting data using only two sequence lengths and then using hypothesis testing to determine if data taken at other sequence lengths are consistent with the hypothesis that the noise is static and Markovian.

ACKNOWLEDGMENTS

RH and STF were supported by the Australian Research Council through the Centre of Excellence in Engineered Quantum Systems CE170100009. This research was supported by the US Army Research Office through grant numbers W911NF-14-1-0098 and W911NF-14-1-0103. IH and JJW gratefully acknowledge contributions from the Canada First Research Excellence Fund, Industry Canada, the Province of Ontario, and Quantum Benchmark Inc.

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

[2] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, "Randomized benchmarking of quantum gates," Phys. Rev. A **77**, 012307 (2008), arXiv:0707.0963.

[3] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, "Scalable and Robust Randomized Benchmarking of Quantum Processes," Physical Review Letters **106**, 180504 (2011), arXiv:1009.3639.

[4] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, "Characterizing quantum gates via randomized benchmarking," Physical Review A **85**, 042311 (2012), arXiv:1109.6887.

[5] 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 Matthias Steffen, "Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking," Physical Review Letters **109**, 080505 (2012), arXiv:1203.4550.

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

[7] Joel J. Wallman, Christopher Granade, Robin Harper, and Steven T. Flammia, "Estimating the Coherence of Noise," New Journal of Physics **17**, 113020 (2015), arXiv:1503.07865.

[8] Joel J. Wallman, Marie Barnhill, and Joseph Emerson, "Robust Characterization of Loss Rates," Physical Review Letters **115**, 060501 (2015), arXiv:1510.01272.

[9] 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), arXiv:1504.06597.

[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), arXiv:1510.02720.

[11] Robin Harper and Steven T. Flammia, "Estimating the fidelity of T gates using standard interleaved randomized benchmarking," Quantum Science and Technology **2**, 015008 (2017), arXiv:1608.02943.

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

[13] Joel J. Wallman, Marie Barnhill, and Joseph Emerson, "Robust characterization of leakage errors," New Journal of Physics **18**, 043021 (2016), arXiv:1412.4126.

[14] C. H. Yang, K. W. Chan, R. Harper, W. Huang, T. Evans, J. C. C. Hwang, B. Hensen, A. Laucht, T. Tanttu, F. E. Hudson, S. T. Flammia, K. M. Itoh, A. Morello, S. D. Bartlett, and A. S. Dzurak, "Silicon qubit fidelities approaching stochastic noise limits via pulse optimisation," ArXiv e-prints (2018), arXiv:1807.09500.

[15] Joel J. Wallman and Steven T. Flammia, "Randomized benchmarking with confidence," New Journal of Physics **16**, 103032 (2014), arXiv:1404.6025.

[16] Jonas Helsen, Joel J. Wallman, Steven T. Flammia, and Stephanie Wehner, "Multi-qubit Randomized Benchmarking Using Few Samples," (2017), arXiv:1701.04299.

[17] Christopher Granade, Christopher Ferrie, and David G. Cory, "Accelerated Randomized Benchmarking," New Journal of Physics **17**, 013042 (2014), arXiv:1404.5275v1.

[18] Christopher Granade, Christopher Ferrie, Ian Hincks, Steven Casagrande, Thomas Alexander, Jonathan Gross, Michal Kononenko, and Yuval Sanders, "QInfer: Statistical inference software for quantum applications," Quantum **1**, 5 (2017), arXiv:1610.00336.

[19] 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," Journal of Physics: Condensed Matter **27**, 154205 (2015), arXiv:1410.2338.

[20] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, Steven T. Flammia, and A. S. Dzurak, "Nonexponential fidelity decay in randomized benchmarking with low-frequency noise," Physical Review A **92**, 022326 (2015), arXiv:arXiv:1502.05119v2.

[21] Christopher J. Wood and Jay M. Gambetta, "Quantification and characterization of leakage errors," Phys. Rev. A **97**, 032306 (2018), arXiv:1704.03081.

[22] Steven T. Flammia and Yi-Kai Liu, "Direct fidelity estimation from few Pauli measurements," Phys. Rev. Lett. **106**, 230501 (2011), arXiv:1104.4695.

[23] B. Dirkse, J. Helsen, and S. Wehner, "Efficient Unitarity Randomized Benchmarking of Few-qubit Clifford Gates," ArXiv e-prints (2018), arXiv:1808.00850 [quant-ph].

[24] D. Katz, J. Baptista, S. P. Azen, and M. C. Pike, "Obtaining confidence intervals for the risk ratio in cohort studies," Biometrics **34**, 469 (1978).

[25] 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), arXiv:1308.2928.

[26] Harrison Ball, Thomas M. Stace, Steven T. Flammia, and Michael J. Biercuk, "Effect of noise correlations on randomized benchmarking," Physical Review A **93**, 022303 (2016), arXiv:1504.05307.