Characterizing errors on qubit operations via iterative randomized benchmarking

Sarah Sheldon, Lev S. Bishop, Easwar Magesan, Stefan Filipp, Jerry M. Chow, and Jay M. Gambetta
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
(Dated: April 27, 2015)

With improved gate calibrations reducing unitary errors, we achieve a benchmarked single-qubit gate fidelity of 99.95% with superconducting qubits in a circuit quantum electrodynamics system. We present a method for distinguishing between unitary and non-unitary errors in quantum gates by interleaving repetitions of a target gate within a randomized benchmarking sequence. The benchmarking fidelity decays quadratically with the number of interleaved gates for unitary errors but linearly for non-unitary, allowing us to separate systematic coherent errors from decoherent effects. With this protocol we show that the fidelity of the gates is not limited by unitary errors, but by another drive-activated source of decoherence such as amplitude fluctuations.

Accurate characterization of control gates is an essential task for developing any quantum computing device. Quantum process tomography (QPT) [1–3] has been the standard method for characterizing quantum gates because, ideally, it produces a full reconstruction of the quantum process. In practice however, QPT suffers from many drawbacks, the most inimical being its exponential scaling in the number of quantum bits (qubits) comprising the system and that it is limited by state preparation and measurement (SPAM) errors. Various methods such as randomized benchmarking (RB) [4–7] and gate set tomography (GST) [8, 9] have recently been developed to help overcome these limitations. RB is both insensitive to SPAM errors and efficient [10]. However, it only extracts a single piece of information, the average gate fidelity. GST on the other hand helps to overcome limitations from SPAM errors by reconstructing an entire library of gates in a self-consistent manner. The price paid for this self-consistent reconstruction is an even worse scaling than QPT.

As control calibration techniques continue to improve and quantum gates approach the fidelity required in fault tolerant quantum computation, it becomes both important and difficult to verify the presence of increasingly small errors. Error verification constitutes a critical first step in a debugging routine since different physical mechanisms can lead to different error types. QPT and GST are often poor choices for error verification since they are time consuming and contain so much information that backing out the presence of specific error types on small scales can be a challenge in itself. In addition, SPAM errors in QPT sets a lower limit on the detectable error strengths [8]. At the other end of the spectrum, while standard RB is efficient the information it contains about the gate is typically not enough to perform any sort of useful error verification. An extension of standard RB, interleaved randomized benchmarking, consists of interleaving a target gate in a benchmarking sequence and provides bounds on the error for the gate of interest [11, 12]. Interleaved benchmarking can identify gates that are poorly calibrated, but does not reveal if the errors are due to decoherence, over-/underrotations, or off-resonance effects amongst other error types. Thus, fast and reliable routines that determine the presence of specific error types are required. Others have proposed to use RB for measuring the unital part of a quantum map [13], correlated errors on a multi-qubit space [14], and recently Ref. [15] has proposed an alternative method for measuring unitary errors. In this paper we propose and experimentally implement a protocol, largely based on the ideas of RB, that verifies the presence of unitary versus non-unitary errors.

A major source of unitary errors in transmon qubits originates from the presence of higher levels, which can be removed by the derivative removal via adiabatic gate (DRAG) protocol [16]. To quantify this error source, we compare experimental randomized benchmarking fidelities for several gate times with two simulations, one assuming a DRAG-corrected pulse shape and the other without DRAG (Fig. 1). The measurements described here are performed on a two-qubit sample consisting of two transmon qubits coupled by a coplanar waveguide resonator, with independent readout resonators for each qubit. The qubit of interest has a transition frequency of 5.0154 GHz and anharmonicity of \(-323\) MHz. \(T_1\) and \(T_2\) are \(45\pm6\) µs and \(53\pm10\) µs, respectively. These characteristic times are the mean values from 500 measurements taken over 14 hours, and the error bars are the standard deviation of this data; each independent experiment is well fit by an exponential decay. The pulses used in the RB sequence are truncated Gaussian pulses having total length equal to four times the standard deviation of the Gaussian and with the DRAG correction applied to the quadrature component.

A typical benchmarking sequence consists of a set of random Clifford gates that together compose to an identity operation [6]. Under realistic assumptions on the noise, the fidelity between the implementation of this sequence with the identity operation decays exponentially as a function of the number of Clifford gates [10]. When the fidelity decay is averaged over many realizations of the random sequence, the decay constant serves as the single metric for the average noise in the system.

The weak anharmonicity, \(\delta\), of the transmon limits the gate fidelity as \(1/\delta\), which can be seen for short gate times in Fig. 1. The experimental data falls below the non-DRAG curve (brown dotted line in Fig. 1), showing that we have partially removed unitary errors due to presence

\(\text{[32x217]}\arXiv{1504.06597v1} \quad \text{[quant-ph]} \quad 24 \text{ Apr 2015}\)
length of the interleaved segment, and the resulting error per segment will thus be linear with $n$. If there are unitary errors of an over-/underrotation type, they will add coherently with $n$, and the fidelity decay will be quadratic to leading order. To see this, suppose we have a single-qubit unitary error of the form

$$U = \exp \left( -i \frac{\epsilon}{2} \hat{r} \cdot \hat{\sigma} \right),$$

where $\epsilon$, $\hat{r}$, and $\hat{\sigma}$ are the error angle, axis of rotation, and vector of Pauli operators respectively. Assuming $\epsilon \ll 1$ we can write $U^n$ to second order in $\epsilon$ as

$$U^n = \mathbb{I} - in\frac{\epsilon}{2} \hat{r} \cdot \hat{\sigma} - \binom{n(2n-1)}{2} \frac{\epsilon^2}{4} (\hat{r} \cdot \hat{\sigma})^2 + O(\epsilon^3).$$

The average fidelity $F$ of the error gate compared to the identity is given by $F = \left( \text{tr}(U^n)^2 + 2 \right) / 6$ and writing $F$ in terms of the benchmarking parameter $\alpha = 2F - 1$ gives

$$\alpha = 1 - \frac{n(2n-1)\epsilon^2}{3},$$

which shows the quadratic dependence in $n$. A similar analysis finds that errors due to a $T_1$ or $T_2$ process do decay linearly in $n$.

We use single sideband (SSB) modulation of our control pulses and calibrate the in-phase/quadrature (IQ) mixers (MITEQ IRM0408LC2Q) for the chosen intermodulation frequency (IF) to ensure only the correct sideband was produced with minimal leakage at the carrier frequency. We then calibrate the in-phase control pulse amplitude and the amplitude of the quadrature component for the DRAG correction. The pulse amplitudes for a $\pi$-pulse ($X_\pi$) and a $\pi/2$-pulse ($X_{\pi/2}$) about the $x$-axis are tuned up by repeating the pulses in the sequence $X_{\pi/2} - (X_{\pi/2})^2$ in order to amplify the errors. The evolution of the qubit’s Bloch vector during the first three points of this sequence is depicted in Fig. 2(a).

We correct for over- or under-rotations by fitting to the measured population of the qubit ground state, $P(\lnot)$ [see Fig. 2b)]. Under the assumption that the error is only an over- or underrotation, it is simple to derive a fitting formula for the amplitude calibration sequences. The fit function for the $X_{\pi/2}$ pulse in this sequence is

$$P(\lnot) = a + \left( \frac{1}{2} (-1)^n \cos(\pi/2 + 2n\epsilon) \right),$$

where $a$ is left as a fit parameter and goes to $1/2$ for perfect $X_{\pi/2}$ pulses. For $X_{\pi}$ the fit function is

$$P(\lnot) = a + \left( \frac{1}{2} \cos(\pi/2 + 2n\epsilon) \right).$$

The angle error, $\epsilon$, found by this fit corresponds to a gate error $r \approx \epsilon^2/6$. After fitting the error, we update the pulse amplitude accordingly.
quadratically and are clearly distinguishable from gates with \( \pi/\lambda \) overrotation, and \( \pi/256 \) overrotation for linear and quadratic model functions.

We performed this protocol with a 16.7\,ns gate time \( T \) and interleaved target sequences. The model with the best fit to the data will have \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]

with \( C_{\text{min}} \) the smallest AIC value for the set of models. The model with the best fit to the data will have \( P_i = 1 \). We calculate the relative probabilities for the three models for iterative randomized benchmarking data with \( X_{\pi/2} \) pulses with no overrotation, \( \pi/128 \) and \( \pi/256 \) overrotations. As detailed in Table II, the calibrated gate with no added error is best fit by a linear model, as expected when there is little unitary error present. The gate with \( \pi/256 \) overrotation is fit best by the combined model. The preferred model according to the AIC for the gate with \( \pi/128 \) error is the quadratic model, but this is in part due to the penalty placed on adding extra parameters to the fit function.

For \( n \) data points and \( k \) fitting parameters, the AIC is given by

\[
C = n \ln \left( \frac{R}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1},
\]

where \( R \) is the residual sum of squares for the fit. The final term in this expression is a correction under the condition that \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]

with \( C_{\text{min}} \) the smallest AIC value for the set of models. The model with the best fit to the data will have \( P_i = 1 \). We calculate the relative probabilities for the three models for iterative randomized benchmarking data with \( X_{\pi/2} \) pulses with no overrotation, \( \pi/128 \) and \( \pi/256 \) overrotations. As detailed in Table II, the calibrated gate with no added error is best fit by a linear model, as expected when there is little unitary error present. The gate with \( \pi/256 \) overrotation is fit best by the combined model. The preferred model according to the AIC for the gate with \( \pi/128 \) error is the quadratic model, but this is in part due to the penalty placed on adding extra parameters to the fit function.

For \( n \) data points and \( k \) fitting parameters, the AIC is given by

\[
C = n \ln \left( \frac{R}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1},
\]

where \( R \) is the residual sum of squares for the fit. The final term in this expression is a correction under the condition that \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]

with \( C_{\text{min}} \) the smallest AIC value for the set of models. The model with the best fit to the data will have \( P_i = 1 \). We calculate the relative probabilities for the three models for iterative randomized benchmarking data with \( X_{\pi/2} \) pulses with no overrotation, \( \pi/128 \) and \( \pi/256 \) overrotations. As detailed in Table II, the calibrated gate with no added error is best fit by a linear model, as expected when there is little unitary error present. The gate with \( \pi/256 \) overrotation is fit best by the combined model. The preferred model according to the AIC for the gate with \( \pi/128 \) error is the quadratic model, but this is in part due to the penalty placed on adding extra parameters to the fit function.

For \( n \) data points and \( k \) fitting parameters, the AIC is given by

\[
C = n \ln \left( \frac{R}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1},
\]

where \( R \) is the residual sum of squares for the fit. The final term in this expression is a correction under the condition that \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]

with \( C_{\text{min}} \) the smallest AIC value for the set of models. The model with the best fit to the data will have \( P_i = 1 \). We calculate the relative probabilities for the three models for iterative randomized benchmarking data with \( X_{\pi/2} \) pulses with no overrotation, \( \pi/128 \) and \( \pi/256 \) overrotations. As detailed in Table II, the calibrated gate with no added error is best fit by a linear model, as expected when there is little unitary error present. The gate with \( \pi/256 \) overrotation is fit best by the combined model. The preferred model according to the AIC for the gate with \( \pi/128 \) error is the quadratic model, but this is in part due to the penalty placed on adding extra parameters to the fit function.

For \( n \) data points and \( k \) fitting parameters, the AIC is given by

\[
C = n \ln \left( \frac{R}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1},
\]

where \( R \) is the residual sum of squares for the fit. The final term in this expression is a correction under the condition that \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]

with \( C_{\text{min}} \) the smallest AIC value for the set of models. The model with the best fit to the data will have \( P_i = 1 \). We calculate the relative probabilities for the three models for iterative randomized benchmarking data with \( X_{\pi/2} \) pulses with no overrotation, \( \pi/128 \) and \( \pi/256 \) overrotations. As detailed in Table II, the calibrated gate with no added error is best fit by a linear model, as expected when there is little unitary error present. The gate with \( \pi/256 \) overrotation is fit best by the combined model. The preferred model according to the AIC for the gate with \( \pi/128 \) error is the quadratic model, but this is in part due to the penalty placed on adding extra parameters to the fit function.

For \( n \) data points and \( k \) fitting parameters, the AIC is given by

\[
C = n \ln \left( \frac{R}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1},
\]

where \( R \) is the residual sum of squares for the fit. The final term in this expression is a correction under the condition that \( n < 40k \). This correction increases the penalty for overfitting when the sample size is small. We compute the \( C \) for three models: linear, quadratic with no linear component, and combined linear and quadratic (see Table II). The relative probability that the \( i \)th model is correct is

\[
P_i = \exp \left[ \frac{1}{2} (C_{\text{min}} - C_i) \right],
\]
explain the deviation of the experiment from the simulated RB [dashed black in Fig. 1(a)]. We conclude that there is an additional source of decoherence that is present under the continuous-driving conditions of an RB experiment. One possible form for such non-unitary error, would be a dephasing proportional to the Rabi rate of the drive, as would result from amplitude fluctuations in the local oscillator, an amplifier, or other microwave electronics along the control line. Simulated RB in presence of such noise (green dot-dashed) shows reasonable agreement with the experimental data. Drive noise with a $1/f$ dependence has been measured in flux qubits [20], and such low frequency noise has been studied in the context of randomized benchmarking [21, 22].

We notice that there is still a deviation from the best fit at the shortest gate time in Fig. 1(a). To understand the origin of this larger error rate we calibrate gates of length 10 ns and apply IRB. For interleaved $I$, $X_{\pi/2}$, and $Y_{\pi/2}$ the iterative benchmarking data appears to decay linearly [Fig. 3(b)]. First, we notice that the error of a $Y_{\pi/2}$ gate is larger than the $X_{\pi/2}$ gate error. We attribute this to our calibration procedure, in which the amplitude of the $Y_{\pi/2}$ is assumed to be equal to the $X_{\pi/2}$ pulse amplitude, but sampling errors in the pulse generation are not taken into account. Second, when the interleaved sequence is $X_{\pi/2}Y_{\pi/2}$ (black stars) a larger decay is observed. This cannot be accounted for by multiplying (dashed black stars) the individual benchmarking parameters, $\alpha$, for the $X_{\pi/2}$ (red circles) and $Y_{\pi/2}$ (magenta diamonds) implying an additional error on the $X_{\pi/2}Y_{\pi/2}$ gate. (Note that, in contrast, no additional error for the $X_{\pi/2}Y_{\pi/2}$ sequence is observed for the 16.67 ns gate, for which the product of $X_{\pi/2}$ and $Y_{\pi/2}$ matches the error for $X_{\pi/2}Y_{\pi/2}$.) The $X_{\pi/2}Y_{\pi/2}$ is not directly calibrated, and the presence of unitary errors here indicates a phase error, despite the fact that SSB modulation ensures the orthogonality of X and Y pulses by imposing a $\pi/2$ phase shift on the IF signal.

After identifying the phase error, we have developed an error amplification sequence similar to those of Fig. 2 in order to quantify an X-Y axes error. The sequence is a repetition of $X_{\pi}Y_{\pi}$ within a Ramsey experiment:

$$X_{\pi/2} - (X_{\pi} - Y_{\pi})^n - Y_{-\pi/2}.$$ 

The fit function for the error case when $X$ and $Y$ are not orthogonal is the same function as for a $\pi/2$ amplitude error given in Eq. 2. The gate error measured by this sequence is $2e^2/3$.

We measure this error as a function of the buffer time between pulses for three different pulse lengths, as shown in Fig. 4. The IRB data was taken with a 3.33 ns buffer indicated by the vertical line [with pulse length of 13.33 ns for the data in Fig. 3(a) and 6.67 ns for Fig. 3(b)]. The gate error is $2 \times 10^{-5}$ for the pulse length corresponding to the 16.67 ns gate, and $3 \times 10^{-5}$ for the 10 ns gate. This is consistent with the IRB data that demonstrates an axis error is present for the 6.67 ns pulse (red squares in Fig. 3 but is not detected for 13.33 ns (violet triangles). The gate error decreases as the buffer time is increased until it levels off by around 15 ns, at which point the resolution of the fit is not better than $1 \times 10^{-5}$. Because the error decreases with longer buffer time, it is likely due to distortions that cause successive pulses to overlap when the time between them is insufficient. Note that this effect is not typically considered in RB, in which it is assumed a pulse knows no history of previous pulses in the sequence. This pulse distortion may be alleviated by further pulse shaping (as shown in [23] with pulse distortions on flux qubits) and will be the subject of future investigations.

We have introduced a variation of randomized benchmarking, useful for distinguishing non-unitary from unitary errors, and have validated this method on a superconducting qubit experiment. IRB will work for most physical unitaries without knowledge of the type of error present. Once a unitary error is discovered, one can de-
We now seek to identify sources of drive-activated non-unitary errors (beyond $T_1$ and $T_2$) that must be limiting our fidelity at this time.

We acknowledge discussions and contributions from Oliver E. Dial, Matthias Steffen, George A. Keefe, and Mary B. Rothwell. This work was supported by ARO under contract W911NF-14-1-0124.

[1] I. L. Chuang and M. A. Nielsen, Journal of Modern Optics 44, 2455 (1997).
[2] J. F. Poyatos, J. I. Cirac, and P. Zoller, Phys. Rev. Lett. 78, 390 (1997).
[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, Phys. Rev. Lett. 102, 090502 (2009).
[4] J. Emerson, R. Alicki, and K. Zyczkowski, Journal of Optics B: Quantum and Semiclassical Optics 7, S347 (2005).
[5] E. Knill, D. Leibfried, R. Reiche, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Phys. Rev. A 77, 012307 (2008).
[6] E. Magesan, J. M. Gambetta, and J. Emerson, Phys. Rev. Lett. 106, 180504 (2011).
[7] R. Barends, J. Kelly, A. Veitia, A. Megrant, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, I.-C. Hoi, E. Jeffrey, C. Neill, P. J. J. O’Malley, J. Mutus, C. Quintana, P. Roushan, D. Sank, J. Wenner, T. C. White, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, Phys. Rev. A 90, 030303 (2014).
[8] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Corcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Phys. Rev. A 87, 062119 (2013).
[9] R. Blume-Kohout, J. King Gambble, E. Nielsen, J. Mizrahi, J. D. Sterk, and P. Maunz, ArXiv e-prints (2013), arXiv:1310.4492 [quant-ph].
[10] E. Magesan, J. M. Gambetta, and J. Emerson, Phys. Rev. A 85, 042311 (2012).
[11] 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, Phys. Rev. Lett. 109, 080505 (2012).
[12] 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, Phys. Rev. Lett. 108, 260503 (2012).
[13] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, ArXiv e-prints (2015), arXiv:1503.07865 [quant-ph].
[14] J. M. Gambetta, A. D. Corcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. 109, 240504 (2012).
[15] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Okhi, Phys. Rev. X 4, 011050 (2014).
[16] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Phys. Rev. Lett. 103, 110501 (2009).
[17] H. Akaike, Second International Symposium on Information Theory , 267 (1973).
[18] K. P. Burnham and D. R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (Springer-Verlag, 2002).
[19] S. J. van Enk and R. Blume-Kohout, New Journal of Physics 15, 025024 (2013).
[20] F. Yoshihara, Y. Nakamura, F. Yan, S. Gustavsson, J. Bylander, W. D. Oliver, and J.-S. Tsai, Phys. Rev. B 89, 020503 (2014).
[21] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak, ArXiv e-prints (2015), arXiv:1502.05119 [quant-ph].
[22] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Phys. Rev. A 89, 062321 (2014).
[23] S. Gustavsson, O. Zwier, J. Bylander, F. Yan, F. Yoshihara, Y. Nakamura, T. F. Orlando, and W. D. Oliver, Phys. Rev. Lett. 110, 040502 (2013).