Nonexponential fidelity decay in randomized benchmarking with low-frequency noise

M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak

Phys. Rev. A **92**, 022326 — Published 11 August 2015

DOI: 10.1103/PhysRevA.92.022326

Non-exponential Fidelity Decay in Randomized Benchmarking with Low-Frequency Noise

M. A. Fogarty,[1] M. Veldhorst,[1] R. Harper,[2] C. H. Yang,[1] S. D. Bartlett,[2] S. T. Flammia,[2] and A. S. Dzurak[1]

[1]*Centre for Quantum Computation and Communication Technology, School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia.*

[2]*Centre for Engineered Quantum Systems, School of Physics, The University of Sydney, Sydney, NSW 2006, Australia*

We show that non-exponential fidelity decays in randomized benchmarking experiments on quantum dot qubits are consistent with numerical simulations that incorporate low-frequency noise and correspond to a control fidelity that varies slowly with time. By expanding standard randomized benchmarking analysis to this experimental regime, we find that such non-exponential decays are better modeled by multiple exponential decay rates, leading to an *instantaneous control fidelity* for isotopically-purified-silicon MOS quantum dot qubits which is 98.9% when the low-frequency noise causes large detuning, but can be as high as 99.9% when the qubit is driven on resonance and system calibrations are favourable. These advances in qubit characterization and validation methods underpin the considerable prospects for silicon as a qubit platform for fault-tolerant quantum computation.

Randomized benchmarking experiments [1, 2] quantify the accuracy of quantum gates by estimating the average decay in control fidelity as a function of the number of operations applied to a qubit. Benchmarking enjoys several advantages over the traditional methods of characterizing gate fidelity that involve quantum process tomography [3, 4], namely that it is insensitive to state preparation and measurement (SPAM) errors, and scales efficiently with the system size. As such, benchmarking protocols (see Figure 1) have become a standard against which different qubit technologies and architectures are compared. Benchmarking experiments have been performed in many different technologies, including trapped ions [1, 5, 6], superconducting qubits [7–9], nuclear magnetic resonance architectures [10], nitrogen-vacancy centers in diamond [11], semiconductor quantum dots in silicon [12], and phosphorous atoms in silicon [13]. Most experiments are fitted using an exponential fidelity decay, which is in line with original theoretical predictions [1, 14], and consistent with the assumption of weak correlation between noise on the gates that is important for accurate fidelity estimates.

When the assumptions of randomized benchmarking are violated, there is no guarantee of observing the characteristic exponential decay curves determined by the average fidelity. This has been noted before in NMR experiments due to spatial inhomogeneity across the sample [10] as well as in numerical simulations [15] of $1/f$ noise and leakage to states outside of the computational subspace. Recent experimental results in spin-based silicon metal-oxide-semiconductor (Si-MOS) quantum dot qubits [12] have also shown non-exponential fidelity decay, and here we directly apply our theoretical modelling to these experiments, but our conclusions are widely applicable.

Here we argue that non-exponential fidelity decay in this semiconductor qubit is indicative of a dephasing-limited decay caused by non-Markovian noise. We first propose a numerical simulation method that incorporates time-dependent effects, primarily a drift in frequency detuning. This detuning drift and other time-dependent low-frequency noise sources lead to decay curves that are effectively integrated over an ensemble of experimental results, each with slightly different "instantaneous" average fidelities, i.e., fidelities that are approximately stable over the course of a single benchmarking run, but that drift over the course of the entire sequence of experiments. These simulations show good qualitative agreement with the observed non-exponential decay from the experiments on isotopically-purified silicon quantum dot qubits [12]. We then give a more quantitative analysis that compares two very simple models that both give good fits to the data: the first is a simplified version of the drift model that postulates that each experimental run has one of only two possible average fidelities; the second model attributes the non-exponential decay to fluctuating SPAM errors. Both of these models have only one additional parameter over the standard benchmarking model, but our quantitative likelihood analysis shows that the simplified drift model is much more probable.

The conclusion of this analysis for the SiMOS quantum-dot qubit is that, while the total average fidelity over a long series of benchmarking runs is 99.6% [12], the instantaneous fidelity can be as high as 99.9% or more when naturally fluctuating environmental noise sources and system calibrations are most favourable. We emphasise that consistent high fidelities such as these may be within reach: further improvements in the system calibration, such as more frequent and accurate estimates of the qubit detuning, could allow these high fidelities directly by exploiting the low-frequency character of the noise. Achieving such high fidelities for single-qubit gate operations gives optimism for exceeding the demanding error thresholds for fault-tolerant quantum computation.

I. BENCHMARKING REVIEW

The standard randomized benchmarking procedure involves subjecting a quantum system to long sequences of randomly sampled Clifford gates followed by an inversion step and a measurement, as depicted in Figure 1. The unitary operations of the Clifford group G are those that map the set of Pauli operators to itself under conjugation. They are a discrete set of gates that exactly reproduce the uniform average gate fidelity, averaged over the set of all input pure states [16].

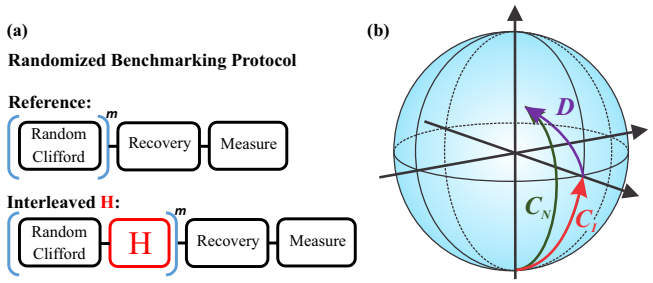

FIG. 1. **a**) Randomized benchmarking consists of applying multiple sequences of random Clifford gates, a final recovery Clifford to ensure that each sequence ends with the qubit in an eigenstate, and reading out the qubit state. In interleaved randomized benchmarking, an additional test-Clifford gate is inserted in between the random Cliffords. **b**) Bloch sphere representation for the breakdown of a general noisy operation $\mathcal{C}_N$ into an ideal $\mathcal{C}_I$ rotation followed by a noise operation $\mathcal{D}$.

An alternate version known as interleaved benchmarking [8] inserts a systematic application of a given gate, such as the H gate shown in Figure 1. The difference from the reference sequence gives information about the specific average gate fidelity of the given gate, rather than the average fidelity additionally averaged over the ensemble of gates.

Consider a general noise process $\mathcal{D}$, depicted in Figure 1, which represents the deviation of a noisy Clifford gate $\mathcal{C}_N$ from an ideal unitary Clifford operation $\mathcal{C}_I$:

$$\mathcal{C}_N = \mathcal{D}\mathcal{C}_I\,.$$

We note that the above equation uses the formalism of completely positive maps [17], and the multiplication corresponds to composition of maps. The standard approach to randomized benchmarking makes the assumption that $\mathcal{D}$ does not depend on the choice of $\mathcal{C}_I$ or other details such as time, but our simulations and of course real experiments will include such a dependence.

The fundamental result of randomized benchmarking [2] is that for sufficiently well-behaved noise the observed fidelities only depend on the average error operation $\mathcal{E}_\mathcal{D}$ averaged over the Clifford group G given by

$$\mathcal{E}_\mathcal{D} = \frac{1}{|\mathsf{G}|}\sum_{\mathcal{C}_I\in\mathsf{G}} \mathcal{C}_I\mathcal{D}\mathcal{C}_I^{-1}\,,$$

as well as any SPAM errors present in the system. Furthermore, standard tools from representation theory reduce this average error operation to one that is nearly independent of $\mathcal{D}$, and is characterized by just a single parameter p. In particular, it is a depolarizing channel $\mathcal{E}$ with $p = p(\mathcal{D})$ being the polarization parameter (i.e., the probability of the information remaining uncorrupted as it passes through the channel). For a d-dimensional quantum system, the action of the depolarizing channel is given by

$$\mathcal{E}(\rho) = p\rho + (1-p)\frac{\mathbb{1}}{d}\,,$$

and the polarization parameter is related to the noisy deviation $\mathcal{D}$ by the average gate fidelity $\bar{\mathcal{F}}_{\text{avg}}(\mathcal{D})$ according to [2]

$$\bar{\mathcal{F}}_{\text{avg}}(\mathcal{D}) = \int \mathrm{d}\psi\langle\psi|\mathcal{D}(|\psi\rangle\langle\psi|)|\psi\rangle = p + \frac{1-p}{d}\,, \qquad (1)$$

where the integral is a uniform average over all pure states.

For a randomized benchmarking sequence comprised of $m+1$ total Clifford gates (including the +1 for the recovery operation), the average sequence fidelity is given by [2]

$$\bar{F}_m = Ap^m + B\,. \qquad (2)$$

Here the parameters A and B quantify the SPAM errors and are given by [2]

$$A = \mathrm{Tr}\big[E\mathcal{D}(\rho - \mathbb{1}/d)\big] \quad , \quad B = \mathrm{Tr}\big[E\mathcal{D}(\mathbb{1}/d)\big]\,,$$

and ρ and E are the noisy state preparations and measurements implemented instead of the ideal desired states and measurements.

A typical benchmarking experiment proceeds by estimating $\bar{F}_m$ for several values of m and fitting to the model in Eq. 2 to extract the p, A, and B fit parameters, and then using Eq. 1 to report an ensemble average of the average gate fidelities $\bar{\mathcal{F}}_{\text{avg}}$ of the gates.

This derivation of Eq. 2 assumes certain features about the noise, namely that it has negligible time and gate dependence, and that non-Markovian effects are not present at timescales on the order of the gate time. The limits to the validity of this assumption have been probed before [15, 18, 19], and in particular it was noted via numerical simulations by Epstein *et al.* [15] that the exponential model of fidelity decay no longer holds in the presence of $1/f$ noise, resulting in a noise floor to the accuracy of the benchmarking experiment.

II. NON-EXPONENTIAL FIDELITY DECAY

A clear deviation from the fidelity decay predicted by Eq. 2 has been observed in a silicon quantum dot qubit [12]. In order to understand the possible origin of this deviation, we have used the qubit characteristics to numerically simulate randomized benchmarking with a realistic noise model. In the experiment, the qubit is defined by the spin state of a single electron. A magnetic field $B_0 = 1.4$ T is applied to create a Zeeman splitting and the qubit is operated using electron spin resonance (ESR) techniques by applying an AC magnetic field with frequency $\omega_0 = \frac{g\mu_B B_0}{\hbar}$. A Rabi π-pulse is realized in $\tau_{op} = 1.6$ µs and using a Ramsey sequence the dephasing time $T_2^* = 120$ µs has been obtained with state-preparation and measurement fidelities of 95% and 92%, respectively [12]. In between consecutive pulses, a waiting time $\tau_w = 0.5$ µs has to be incorporated, due to the operation of the analog microwave source.

The set of Clifford gates is generated using the set $[\pm X, \pm\frac{1}{2}X, \pm Y, \pm\frac{1}{2}Y]$ that are realized using Rabi pulses, and the identity simulated with a waiting time equal to a π-pulse.

A non-exponential fidelity decay can be caused by leakage, where population in the two qubit states is lost to other levels [15, 20, 21]. For example, in Ref. [15] it has been shown that in the presence of a third level, the sequence fidelity for large m could approach 1/3, instead of 1/2 (although this benchmarking protocol used a different gateset than the standard one). While leakage is an important aspect in multi-dot qubits, such as singlet-triplet qubits or exchange-only qubits that possess accessible non-qubit spin states, a qubit encoded in a spin-1/2 system is inherently two-dimensional. Higher energy levels of the quantum dot, or valley degeneracies, represent different degrees of freedom, rather than leakage channels. Leakage can occur through loss of the electron, but we note that the qubit system experiences a T_1 time on the order of seconds and we have measured the absence of tunnelling during a sequence. As further evidence of negligible leakage, we note that experimentally we observe the spin-up and down fractions are symmetric around the half-visibility-plus-offset, as observed in Figure 4b and c.

A. Non-Markovian noise in a quantum dot qubit

Within the experiment, Ramsey sequences have been performed in between benchmarking sequences to recalibrate the resonance frequency of the qubit and to compensate drifts due to, for example, the superconducting magnet. These drifts, in combination with errors in setting the resonance frequency, cumulate in a time dependency within the system and result in an apparent T_2^* for the randomized benchmarking experiment. This decoherence time is also dependent upon the duration of the data acquisition. The non-Markovian noise processes that are expected to determine T_2^* can be modelled as a random walk of the detuning $\Delta\omega$ away from the ideal operation frequency ω_0, over timescales greater than a single run of a random Clifford sequence. In order to simulate an ensemble of results, the $\Delta\omega$ term is selected randomly from a Gaussian distribution of normalized variance:

$$\sigma_{\text{op}} = \frac{\tau_{\text{op}}}{2\pi\sqrt{2\ln(2)}T_2^*}\,.$$

Using this distribution, we have numerically simulated benchmarking experiments and the results are shown in Fig. 2. In this simulation we have approximated the timescale of the low frequency noise to be on the order of a single run, and as such the detuning is constant over a single trace. However, between each trace the detuning is sampled randomly from a distribution as shown in Fig.2b and c. The individual traces correspond to a given detuning $\Delta\omega$ and result in the "instantaneous" fidelity of the qubit. While the individual traces are decaying exponentially, the averaged fidelity (bold blue) obtained from the Gaussian ensemble is clearly non-exponential. We have also included the case of a Lorentzian distribution of detunings (red), resulting in a non-exponential decay as well. In the simulation, the only error source is dephasing, whereas in the experiment, other errors might be present such as pulse-errors. Inclusion of such errors will still result in non-exponential decays, provided dephasing is a significant source of error.

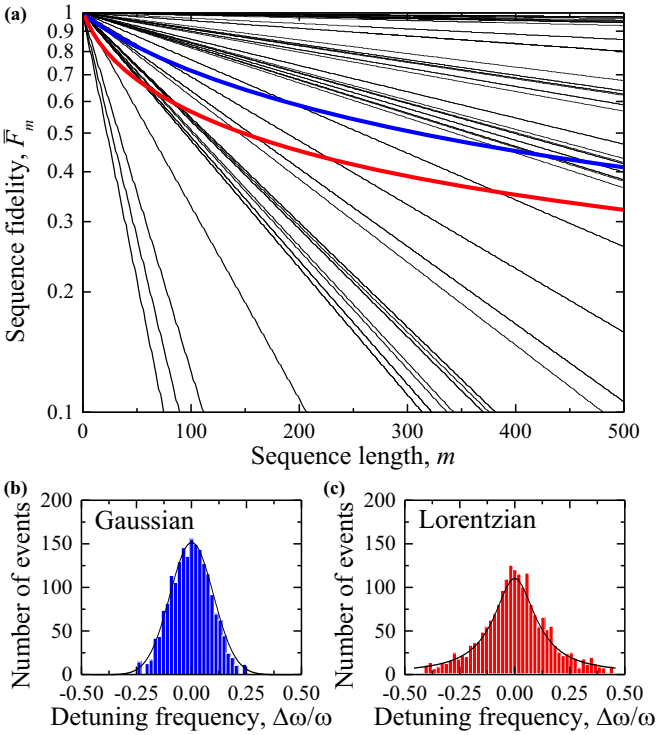

FIG. 2. **a**) Sequence fidelity as a function of sequence length m, with the qubit subject to Gaussian distributed T_2^* associated noise. Each black line represents a fidelity decay for one particular value of detuning $\Delta\omega$ (only 10% of traces shown). The linear decay on the logarithmic scale illustrates that these individual traces are indeed exponential while the ensemble average (bold blue) is non-exponential. The bold red line is the ensemble average for a Lorentzian distributed noise. **b**) Gaussian distributed detuning frequencies and **c**) Lorentzian distributed detuning frequencies associated with individual traces.

As low-frequency drift of the qubit resonance frequency can lead to non-exponential fidelity decay, we hypothesize that some ensemble of experiments with varying decay rates is the correct explanation for the non-exponential behavior of the experimental benchmarking data [12]. To support this hypothesis, we use the Akaike information criterion to show that a simple model allowing for differing fidelity rates better explains the data than an alternative explanation that assumes fluctuating SPAM errors in the standard (zeroth order) model.

B. Eliminating the constant for a single-qubit randomized benchmarking model

The parameters A and B in Eq. 2 are nuisance parameters that do not convey information about the desired control fidelity. Eliminating one of these parameters, in this case B, will further constrain the zero order model and allows deviations to be more clearly identifiable. A further advantage of removing the parameter B is to allow fitting of a linear function on a log-linear plot where deviation from standard

assumptions of randomized benchmarking will show clearly as a non-exponential decay trace. In Ref. [12] the randomized benchmarking protocol was modified to eliminate B from the zero order model. We first provide a theoretical justification for this approach that conforms with the standard assumptions of randomized benchmarking. We note that this approach applies only to qubits ($d = 2$), and demonstrate the deviation of the measured data from the expected exponential is highlighted via this method. Recall that the zero-order model fits the average fidelity of a gate sequence to a simple formula as follows [2]:

$$\bar{F}_m^{\uparrow} = A^{\uparrow} p^m + B^{\uparrow}\,, \tag{3}$$

where the qubit is initialized as $|\uparrow\rangle\langle\uparrow|$, the final gate in the random benchmarking sequence is chosen to return the state to $|\uparrow\rangle\langle\uparrow|$, and $\bar{F}_m^{\uparrow}$ is the survival probability of this state. To eliminate the constant $B^{\uparrow}$ from this sequence, we perform a second set of similar randomized sequences, with the difference being that the final $(m+1)^{\text{th}}$ gate is set to change the state to $|\downarrow\rangle\langle\downarrow|$. For these runs, we consider the survival probability for yielding the measurement outcome $E^{\downarrow}$, where in the ideal case the final state $\rho = E^{\downarrow} = |\downarrow\rangle\langle\downarrow|$. This is the survival probability for each run $\bar{F}_m^{\downarrow}$. Under the same assumptions we have

$$\bar{F}_m^{\downarrow} = A^{\downarrow} p^m + B^{\downarrow}\,. \tag{4}$$

Combining these two equations by defining $\tilde{F}_m \equiv \bar{F}_m^{\uparrow} - (1 - \bar{F}_m^{\downarrow})$, we have:

$$\tilde{F}_m = \tilde{A} p^m + (B^{\uparrow} + B^{\downarrow}) - 1\,, \tag{5}$$

where $\tilde{A} = A^{\uparrow} + A^{\downarrow}$.

Recall that $B^{\uparrow} = \mathrm{Tr}\big[E^{\uparrow}\mathcal{D}(\mathbb{1}/d)\big]$, where $\mathcal{D}$ is the average noise operator. For the $\bar{F}_m^{\downarrow}$ runs, the derivation is identical, apart from the final change to the $|\downarrow\rangle\langle\downarrow|$ state using a π pulse $\mathcal{X}$, so we have $B^{\downarrow} = \mathrm{Tr}\big[E^{\downarrow}\mathcal{D}(\mathcal{X}(\mathbb{1}/d))\big]$. One expects that $\mathcal{X}$ is close to unital, i.e., $\mathcal{X}(\mathbb{1}/d) \simeq \mathbb{1}/d$. Under the assumption that this is true (an assumption that will be respected up to violations no larger than the gate infidelity), and noting that $E^{\uparrow} + E^{\downarrow} = \mathbb{1}$ for qubits ($d = 2$) and that $\mathcal{D}$ is trace-preserving, $B^{\downarrow}$ can be re-expressed as follows:

$$\begin{aligned} B^{\downarrow} &= \mathrm{Tr}\big[E^{\downarrow}\mathcal{D}(\mathbb{1}/2)\big] = \mathrm{Tr}\big[(\mathbb{1} - E^{\uparrow})\mathcal{D}(\mathbb{1}/2)\big] \\ &= \mathrm{Tr}\big[\mathcal{D}(\mathbb{1}/2)\big] - \mathrm{Tr}\big[E^{\uparrow}\mathcal{D}(\mathbb{1}/2)\big] = 1 - B^{\uparrow}\,. \end{aligned} \tag{6}$$

Therefore by subtracting the average results of the data-set $(1 - \bar{F}_m^{\downarrow})$ from the average results of the data-set $\bar{F}_m^{\uparrow}$ we can obtain a data set that is distributed according to the model

$$\tilde{F}_m = \tilde{A} p^m \tag{7}$$

under the standard benchmarking assumptions on the noise.

The data from Ref. [12] consist of 8 data sets (one reference set and 7 interleaved sets). The experiment, in order of operation, comprised of measuring 50 single shots which were randomly distributed over $\bar{F}_m^{\uparrow}$ and $\bar{F}_m^{\downarrow}$. The sequence randomization protocol was carried out 10 times, first for the reference and then all 7 interleaved sets. For each data set, the gate sequences of the lengths $m \in \{2, 3, 5, 8, 13, 21, 30, 40, 50, 70, 100, 150\}$ were measured. This entire process was then repeated 50 times for a total of 2,400,000 single shot measurements. The fast Ramsey recalibrations were performed at approximately 10 minute intervals.

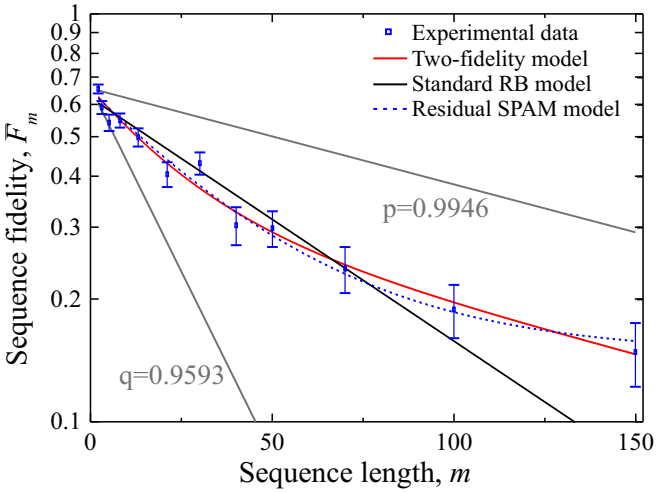

FIG. 3. Semi-log plot of $F_m^{\uparrow} - (1 - F_m^{\downarrow})$ for the reference sequence of randomized benchmarking on a silicon quantum dot qubit [12]. Both the two fidelity model and a single fidelity model including residual SPAM can fit the data, but for the single fidelity model an unreasonably large SPAM has to be included.

This amount of randomization is at least an order of magnitude more than in previous experiments [1, 5, 7–11, 13]. Each randomized protocol was performed 50 times in order to estimate the survival probability.

C. Analysis of experimental data

To quantify the quality of our fits, we require estimates of the variance in the data of Ref. [12]. The observed variance of the data matched to within 5-40% of the theoretical upper bounds derived in Ref. [18] when the gate length was shorter than 20 (so that the $m(1-\bar{\mathcal{F}}_{\text{avg}}) \ll 1$ assumption discussed in that reference was satisfied). Accordingly, the observed experimental variance was used as a reliable estimate of the actual variance of the distribution. It should be noted that the observed variance actually decreased for gate lengths of 100 or greater. One explanation for this unexpected behaviour is that some of the sequences become saturated to something close to a completely mixed state before reaching those sequence lengths.

Figure 3 shows the data from the reference dataset plotted on a semi-log plot. The confidence bounds are 95% and the data is clearly non-linear (i.e. the decay is not a simple exponential). Similar deviation from the linear fit was noted in each of the data sets, with the best-fit linear model consistently underestimating $\bar{F}_m$ for $m \geq 100$. Ref.[2] outlines a higher (first) order fitting for the fidelity decay which includes gate dependent noise. With the elimination of the parameter

Dataset	Akaike Information Criteria $\tilde{A}p^m + \tilde{B}$	$\tilde{A}p^m + \tilde{A}q^m$	Comparison
Ref	-16.93	-25.29	65.44
I	-46.19	-57.12	238.10
X	-54.52	-59.99	15.43
X/2	-62.89	-63.79	1.56
-X/2	-57.77	-64.34	26.69
Y	-36.06	-50.43	1317
Y/2	-36.04	-46.39	172.0
-Y/2	-46.37	-63.32	4815

TABLE I. Akaike information criterion for standard and interleaved randomized benchmarking. The comparison column specifies how many times as probable is the $\tilde{A}p^m + \tilde{A}q^m$ model to minimize information loss as compared to the $\tilde{A}p^m + \tilde{B}$ model.

B as outlined above, the inclusion of higher orders results in $\tilde{F}_m^1 = \tilde{A}p^m + \tilde{C}(m-1)(q-p^2)p^{m-2}$. The first order equation did not fit the data significantly better than the zero order equation.

Two other possible explanations are considered. First, it may not be possible to entirely eliminate the constant term (B) due to a violation of one of the assumptions in the above derivation. A second explanation is that low-frequency noise leads to detuning, and hence time-dependent errors on the gates in some of the experiments. The first, which we denote the *residual SPAM model*, can be modelled by reverting to a formula of the form $\tilde{F}_m = \tilde{A}p^m + \tilde{B}$, where now $\tilde{B}$ represents residual SPAM errors that were not eliminated under the assumptions that led to the derivation of Eq. 7. We consider the simplest possible model for the second explanation – the *two fidelity model* – by fitting the fidelity decay to a formula of the form $\tilde{F}_m = \tilde{A}p^m + \tilde{A}q^m$. This represents an attempt to model the data by simplifying the ensemble of experiments by reducing them to just two different *equally weighted sequence behaviours:* one with a high-fidelity rate (related by the usual measure to p) and one with a lower fidelity rate (similarly related to q). This model has fewer parameters than the Gaussian or Lorentzian drift models, and is much easier to fit. In this interpretation, we have successfully eliminated the B parameter as per Eq. 7, but time variation gives us the two different polarization parameters p and q, with the decay rate for each sequence sampled randomly with equal probability. As can be seen in Figure 3, both models fit the data substantially better than the simple exponential of the zero order model.

Although the residual SPAM model produces a good fit to the experimental data, it does so with the equivalent of an unusually large SPAM parameter $\tilde{B}$ of around 0.14 corresponding to an individual fitting of $B^{\uparrow} = 0.56$ and $B^{\downarrow} = 0.58$. This represents in the theoretical model a very large bias in the expectation value of the spin-up measurement on the asymptotic value of the sequence fidelity away from the theoretical value of 0.5, which is not observed in the experiment.

To compare the residual SPAM and two-fidelity models quantitatively, it is possible to calculate the log likelihood and Akaike information criterion [22] for the two models. Because we don't have the actual distribution of the test statistic, we make the assumption that the samples contained in the underlying data are independent and the Gaussian distributed limit is appropriate. This assumption is well-justified as we have a large number of independent data sets. The distribution $\tilde{F}_m$ can therefore be approximated by a Gaussian distribution with a variance estimated by the observed variance at each gate length. The log likelihood of the observed data, given each of the two models, can then be calculated using standard methods, as follows.

The Akaike information criteria used was $2 * (\text{Number of parameters} - \text{loglikehood})$. The log likelihood is calculated as:

$$\sum_{s=\substack{\text{sequence}\\ \text{lengths}}} \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma_s^2}}\right) - \frac{1}{2\sigma_s^2}\left[\mu_s - x_s\right]^2 \right) \qquad (8)$$

where σ_s^2, is the variance for sequence s, μ_s is the measured average at that sequence and x_s is the predicted average.

Table I shows the calculated Akaike information criterion for each of the experimental datasets. As can be seen, the two fidelity model better explains the data, significantly so on all but one of the datasets. Although such a model is a simplified version of the drift model, the fact that it fits the data well and is physically motivated supports its adoption as the most likely explanation of the non-exponential curve seen in the data.

D. Interpreting the two fidelity model

Since the two fidelity model is the quantitatively preferred model, a natural question arises: how should we interpret the model parameters? The obvious interpretation of the two parameters p and q is as presented in table II; that their difference represents the characteristic spread of the actual underlying ensemble of fidelities from which the benchmarking data are sampled. Such an interpretation is natural and compelling, however it remains an open problem to quantify such a connection more carefully. In particular, it would be interesting to give a direct connection to a more general drift model, since these are easier to interpret physically, but much harder to fit and analyze statistically.

By considering the non-exponential decay manifesting as the average over an ensemble of results, the fidelity can be considered to be operating under two regimes as depicted in Figure 4a. Firstly, dominating the observed fidelity decay at low m, there is a rapid decay rate dominated by traces of large detuning $\Delta\omega$. Secondly, for large m, these traces of large detuning will fast approach the constant B and so will not influence the decay slope in the fidelity; the fidelity decay rate for large m is then governed by long-lived traces of smaller detuning.

In Figure 4a, each of the data points are an average over 25,000 experimental repetitions as presented by the two accompanying histograms (Figure 4b for $m = 2$ and Figure 4c for $m = 150$). Each histogram separately shows the measured probability, averaged over 50 repetitions, for the spin-up and spin-down observables as expected at the end of a noiseless

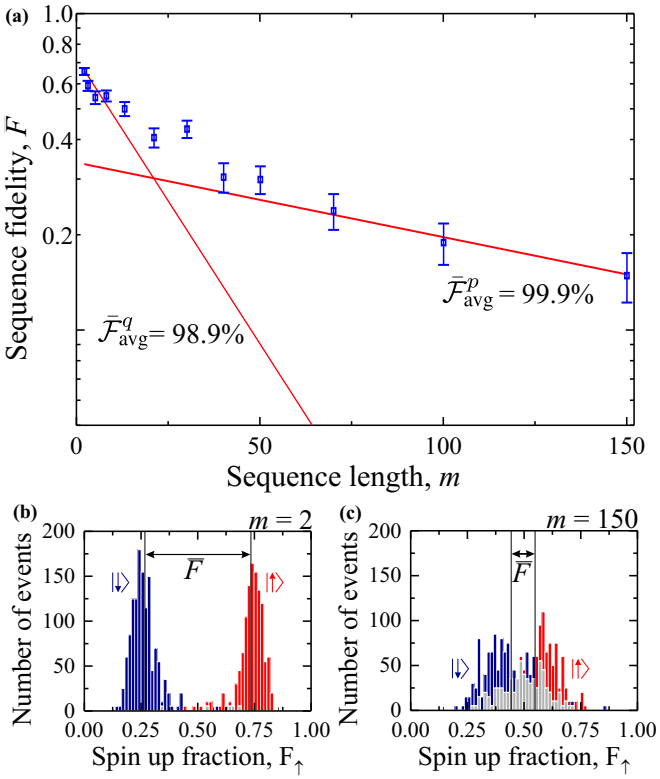

FIG. 4. (**a**) Reference sequence of randomized benchmarking on a silicon quantum dot qubit [12]. The separate fidelities from the two fidelity model have been plotted to show how the initial decay is dominated by the low q value, whereas the higher value of p is indicative of the average decay in the longer lived high fidelity regime. Histogram of spin-up $|\uparrow\rangle$ and spin-down $|\downarrow\rangle$ corresponding with data point $m = 2$ (**b**) and $m = 150$ (**c**). Results with expected spin up outcome are shown in red while blue represents data with expected spin down result. The grey regions illustrate the overlapping areas.

version of the applied random sequence. From Fig. 4 we infer that the instantaneous fidelity approaches a peak fidelity of 99.9% when the microwave frequency is on resonance and that the fidelity can drop to 98.9% in the presence of large detunings. The time-average performance will lead to a fidelity in between, consistent with the 99.6% fidelity quoted in Ref. [12].

A similar analysis can also be applied to the interleaved gate sequences, to extract average fidelities for the individual gates. See table II for p and q values for each set. For each of the interleaved gates, we find a high and a low gate fidelity comparable with the two values quoted for the standard benchmarking scheme above, although the numerical instability in calculating the gate fidelities for interleaved benchmarking leads to larger uncertainties in these values, of the order of 1% for 95% confidence margins. We note that a naive application of the method to derive fidelities for interleaved gate sequences as outlined by Ref. [8] will, for certain gates such as the Y gate, yield a fidelity $\bar{\mathcal{F}}_{\rm avg} > 100\%$. Such results may be an indication of correlated low frequency noise, where some Clifford gates can echo out noise; such an effect would result in decays which are slower than that of the reference set, and

Dataset	p	q
Ref	0.995	0.959
I	0.993	0.946
X	0.993	0.952
X/2	0.993	0.947
-X/2	0.991	0.947
Y	0.993	0.964
Y/2	0.991	0.952
-Y/2	0.990	0.911

Dataset	$\bar{\mathcal{F}}^p_{\rm avg}$	$\bar{\mathcal{F}}^q_{\rm avg}$
Ref	99.9%	98.9%

TABLE II. Calculated p and q values for the two fidelity model. The gate fidelity estimates ($\bar{\mathcal{F}}_{\rm avg}$) reported for the reference run are the high (p) gate fidelity estimate and low (q) gate fidelity with the 95% confidence margins are $\pm 0.06\%$ and $\pm 0.5\%$ respectively. We further note that calculated p and q values will result in an inaccurate interleaved gate fidelity as given by the process outlined by Ref. [8] due to the low-frequency noise.

break the assumptions of interleaved randomized benchmarking. However, the large uncertainty in our estimated fidelities for interleaved gates due to statistical noise does not allow us to definitively test for such an effect with the current dataset.

III. CONCLUSIONS

We have analyzed the non-exponential decay in randomized benchmarking experiments on Si-MOS quantum dot qubits, and found that the most plausible explanation of this decay is drift in detuning frequencies. Our simulation of temporal integration over a spectrum of time-dependent detuning frequencies qualitatively reproduces the observed fidelity decay of previously conducted experiments [12]. In addition, we have quantitatively ruled out a competing model by showing agreement of a simplified ensemble (the two fidelities model) that is much more probable. This yields confidence that detuning drift is the correct explanation for the origin of such a non-exponential fidelity decay.

Low frequency noise leads to a time-varying fidelity that is relatively constant over a given sequence but can vary between sequences. We have therefore defined an "instantaneous fidelity" to characterise the performance of a gate during a single sequence, and we can consider how this instantaneous fidelity varies in time from sequence to sequence. Fitting the randomized benchmarking data with a two-fidelity model demonstrates that silicon MOS quantum dot qubits can already exhibit an "instantaneous" control fidelity of 99.9%. This is achieved when the system is correctly calibrated and the microwave frequency is on resonance. However when the noise causes large detuning the fidelity drops to 98.9%. We anticipate that the higher fidelity can be achieved consistently (i.e., made time-independent) as improvements in the readout fidelity appear feasible and better calibration could be obtained by performing optimized Ramsey protocols to calibrate the resonance frequency for each experiment [23].

These results raise several intriguing questions. The first

is to quantitatively link the simple and easy to analyze two fidelity model to the Gaussian or Lorentzian drift models. Alternatively, directly fitting a drift ensemble to the data would give a better picture of the source of the non-exponential fidelity decay, but this approach risks overfitting, and is already difficult for the simple case of Gaussian-distributed detuning.

Secondly, there are several different models discussed in this work, each of which are capable of analysing different forms of breaches in the standard assumptions of randomized benchmarking. The reduced parameter representation for the fidelity ($\tilde{F}$) not only allows for higher accuracy in fitting, but it is capable of immediately identifying a deviation from the expected result under standard assumptions. Further it is capable of identifying the presence of certain types of noise if $\tilde{F}$ is to asymptote to a non-zero offset.

Finally, there is at least one other natural competing explanation for the non-exponential decay. It might be the case that long benchmarking sequences saturate the exponential decay rates and have slower decay on very long timescales. If this were the case, then fitting to sequences that were "too long" would certainly bias one toward seeing non-exponential decay and reporting fidelities that were higher than warranted by the analysis. Therefore, deriving stopping criteria for the maximum sequence length and deriving tests that rule out this alternate explanation is a further important open question for future work.

ACKNOWLEDGMENTS

We thank Chris Ferrie and Chris Granade for helpful discussions. The authors acknowledge support from the Australian Research Council (CQC2T - CE11E0001017 and EQuS - CE11001013), the NSW Node of the Australian National Fabrication Facility, the US Army Research Office (W911NF-13-1-0024, W911NF-14-1-0098, W911NF-14-1-0103 and W911NF-14-1-0133), and by iARPA via the MQCO program. M.V. also acknowledges support from the Netherlands Organization for Scientific Research (NWO) through a Rubicon Grant. S.T.F. also acknowledges support from an ARC Future Fellowship (FT130101744).

[1] E. Knill *et al.*, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).
[2] E. Magesan, J.M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, Phys. Rev. Lett. **106**, 180504 (2011).
[3] I. Chuang and M. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, J. Mod. Opt. **44**, 2455 (1997).
[4] J.F. Poyatos, J.I. Cirac, and P. Zoller, Complete characterization of a quantum process: the two-bit quantum gate, Phys. Rev. Lett. **78**, 390 (1997).
[5] J.P. Gaebler *et al.*, Randomized benchmarking of multiqubit gates, Phys. Rev. Lett. **108**, 260503 (2012).
[6] T.P. Harty, *et al.*, High-Fidelity Preparation, Gates, Memory, and Readout of a Trapped-Ion Quantum Bit, Phys. Rev. Lett. **113**, 22, 220501 (2014).
[7] J.M. Chow *et al.*, Randomized benchmarking and process tomography for gate errors in a solid-state qubit, Phys. Rev. Lett. **102**, 090502 (2009).
[8] E. Magesan *et al.*, Efficient measurement of quantum gate error by interleaved randomized benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).
[9] R. Barends *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature **508**, 500-503 (2014).
[10] 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).
[11] F. Dolde *et al.*, High-fidelity spin entanglement using optimal control, Nat. Comm. **5**, 3371 (2014).
[12] M. Veldhorst *et al.*, An addressable quantum dot qubit with fault-tolerant control-fidelity, Nat. Nano. **9**, 981 (2014).
[13] J.T. Muhonen *et al.*, Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking, 2015 J. Phys. Cond. Matt. **27** 154205 (2015).
[14] J. Emerson, R. Alicki and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B **7**, S347 (2005).
[15] 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).
[16] 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).
[17] M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, (2000).
[18] J.J. Wallman and S.T. Flammia, Randomized Benchmarking with Confidence, New J. Phys. **16**, 103032 (2014).
[19] C. Granade, C. Ferrie, and D.G. Cory, Accelerated Randomized Benchmarking, New J. Phys. **17**, 013042 (2015).
[20] J.J. Wallman, M. Barnhill, and J. Emerson, Characterization of Leakage Errors via Randomized Benchmarking, arXiv preprint arXiv:1412.4126 (2014).
[21] J. Wallman, C. Grenade, R. Harper, and S. T. Flammia, Estimating the Coherence of Noise, arXiv preprint arXiv:1503.07865 (2015).
[22] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Auto. Control*, **19**, 716 (1974).
[23] M.D. Shulman *et al.* Suppressing qubit dephasing using real-time Hamiltonian estimation, Nat. Comm. **5**, 5156 (2014).