Probing Context-Dependent Errors in Quantum Processors

Kenneth Rudinger,[1,*] Timothy Proctor,[1] Dylan Langharst,[1,2] Mohan Sarovar,[1] Kevin Young,[1] and Robin Blume-Kohout[1]

[1] *Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico 87123, USA, and Livermore, California 94550, USA*
[2] *Behrend College, Pennsylvania State University, Behrend, Erie, Pennsylvania 16563, USA*

(Received 9 November 2018; published 3 June 2019)

Gates in error-prone quantum information processors are often modeled using sets of one- and two-qubit process matrices, the standard model of quantum errors. However, the results of quantum circuits on real processors often depend on additional external "context" variables. Such contexts may include the state of a spectator qubit, the time of data collection, or the temperature of control electronics. In this article, we demonstrate a suite of simple, widely applicable, and statistically rigorous methods for detecting context dependence in quantum-circuit experiments. They can be used on any data that comprise two or more "pools" of measurement results obtained by repeating the same set of quantum circuits in different contexts. These tools may be integrated seamlessly into standard quantum device characterization techniques, like randomized benchmarking or tomography. We experimentally demonstrate these methods by detecting and quantifying crosstalk and drift on the publicly accessible 16-qubit ibmqx3.

DOI: 10.1103/PhysRevX.9.021045 Subject Areas: Quantum Information

I. INTRODUCTION

Quantum characterization, verification, and validation (QCVV) [1–21] tools are methods to probe the *in situ* behavior of quantum information processing hardware. Most QCVV protocols assume a "standard model" of errors in which each imperfect quantum operation is represented by a single, completely positive, trace-preserving linear map on density matrices (i.e., a *process matrix*). Although this model can describe many deviations from ideal behavior, including coherent errors caused by a fixed Hamiltonian and stochastic errors caused by white noise fluctuations, there are many other possible failure modes whose impacts on both quantum error correction (QEC) and near-term quantum information processing applications are not yet well understood. Many of them manifest as a *dependence* of the error process on some external variable, or *context*, that is not supposed to affect qubit behavior [22]. For example, an error rate might drift over time [4,23–25] or increase when a nearby qubit is being measured or driven [7–9,26–28]. These effects are important in their own right. They might contribute significantly to the device's total observed error rate [7–9], and they may have consequences for QEC [26,29–33]. Context dependence is also important because it can interfere with standard QCVV techniques such as randomized benchmarking (RB) [5–21] or gate set tomography (GST) [1–4] and potentially invalidate conclusions drawn from them [25].

In this paper, we propose and demonstrate a practical, statistically rigorous toolkit for detecting whether a quantum circuit's observable behavior depends on external variables. The underlying statistical tasks here are old and well studied [34–37], so we make no claims of statistical novelty. Instead, our focus is on choosing and harnessing established statistical techniques for detecting context dependence in QCVV, using the type of data most often found in quantum device characterization and circuit-based experiments. Almost all such experiments generate *count data*: the aggregated outcomes of N repetitions of one or more quantum circuits that each begin with a state preparation and end with a measurement.

Usually, all the measurement results for a single circuit are collected into a single "pool." This collection precludes testing for variation, because a single pool of counts is always perfectly consistent with a single underlying set of probabilities for the observed outcomes. However, some data have additional structure, such as time stamps, that defines a natural division into two or more pools that are each associated with a different "context." Then, we can look for *significant* variation in the circuit behavior between contexts (Fig. 1). For example, flipping two coins 100 times and getting 49 heads for one coin and 55 for the other is intuitively consistent with the claim that the coins are identically biased; the variation is typical of random finite-sample fluctuations. Observing instead 28 heads for one coin and 72 heads for the other is strong evidence that the coins actually have different biases. We can address this

[*] kmrudin@sandia.gov

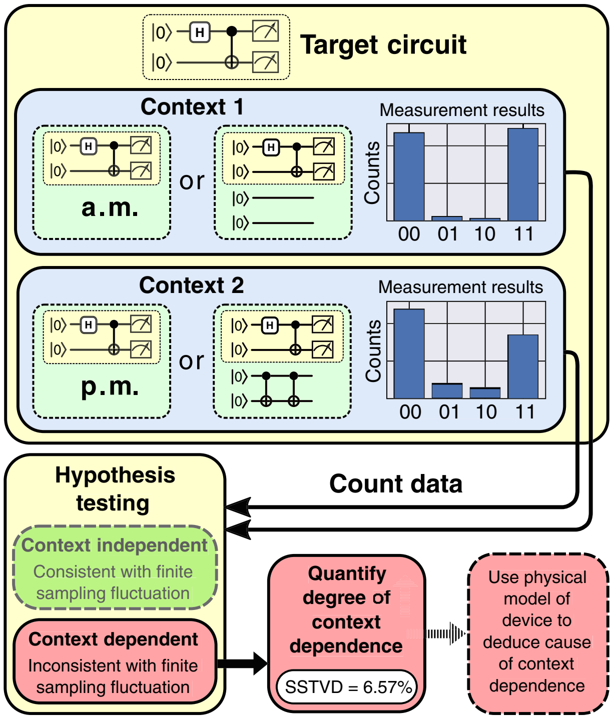
Target circuit
|0⟩
H
Context 1
Measurement results
a.m.
or
Counts
00 01 10 11
Context 2
p.m.
Hypothesis testing
Count data
Context independent
Consistent with finite sampling fluctuation
Context dependent
Inconsistent with finite sampling fluctuation
Quantify degree of context dependence
SSTVD = 6.57%
Use physical model of device to deduce cause of context dependence

FIG. 1. An illustration of how to detect and quantify context dependence in a quantum information processor by repeatedly performing a quantum circuit in two or more contexts. In this simple example, a Bell state is prepared during two different time periods (a.m./p.m.), to test for time variation, or while an adjacent pair of qubits is or is not being driven, to test for crosstalk. The measurement outcome frequencies for the two contexts are compared to determine if the circuit behavior is the same across contexts. If not, the change is quantified. Multiple test circuits and a physical model of the device can sometimes enable identification of the underlying cause and indicate the size of the effect.

question formally using *statistical hypothesis testing*, a standard framework for rigorously deciding if there is sufficient evidence to reject a base assumption, known as a *null hypothesis*. In the tools we propose, our null hypothesis is that there is no context dependence, and we seek statistically significant evidence in the data to the contrary.

This paper is structured as follows. In Sec. II, we present hypothesis-testing techniques for detecting context dependence in count data from one or more circuits. In Sec. III, we adapt these context dependence *detection* tools to the task of context dependence *quantification*. In Sec. IV, we simulate applying these techniques to detect drift, demonstrating that these methods can clearly highlight context-dependent errors. In Sec. V, we apply our techniques to drift and crosstalk detection and quantification on the ibmqx3 [38], a publicly accessible superconducting quantum processor. In Sec. VI, we discuss the relationship between our tools and simultaneous RB [7], a popular crosstalk quantification technique, and we conclude in Sec. VII.

II. DETECTING CONTEXT DEPENDENCE

A. Single-circuit data

First, we consider how to *detect* context dependence in a *single* quantum circuit. Suppose this circuit has $M \geq 2$ possible measurement outcomes, indexed by $m = 1, 2, \ldots, M$. In general, if a circuit has n qubits (and all n qubits are read out at the end of the circuit), then $M = 2^n$. Note that we could also choose to measure only a subset of the qubits in the system or marginalize multiqubit data over some of the qubits. Let this circuit be performed repeatedly in each of C different contexts, indexed $c = 1, 2, \ldots, C$. For example, the contexts might correspond to distinct time intervals or to driving (or not driving) neighboring qubits (see Fig. 1). For each context c, the circuit defines a probability distribution over the possible measurement results:

$$\mathbf{p}_c = (\mathsf{p}_{c,1}, \mathsf{p}_{c,2}, \ldots, \mathsf{p}_{c,M}). \tag{1}$$

These are the probabilities for obtaining each of the M measurement outcomes, *after* averaging over any other unaccounted-for contexts that might vary within a c-indexed context. For example, time is a continuously varying context variable, and a time period context is a coarse-graining over time. Thus, in this example, each $\mathbf{p}_c$ is the probability distribution after this time averaging. An experiment consists of running our circuit N_c times in each context c and recording the total counts for each measurement outcome m. This procedure effectively samples from each of the $\mathbf{p}_c$ distributions, producing measurement results $x = \{\mathbf{x}_c\}$. Here,

$$\mathbf{x}_c = (\mathsf{x}_{c,1}, \mathsf{x}_{c,2}, \ldots, \mathsf{x}_{c,M}) \tag{2}$$

is a vector of positive integers summing to N_c, representing the observed counts from N_c repeats of the circuit in context c. In terms of the data, context independence holds iff all of the data are drawn from the same underlying probability distribution $\mathbf{p}_0$. To detect context dependence, we therefore ask whether the measurement results in different contexts are consistent with being drawn from a single distribution. This is a hypothesis-testing problem: We are looking for evidence to reject the null hypothesis that the underlying distributions are context independent.

In general, hypothesis testing is the following procedure:

(1) Choose a *statistic*. This statistic is a function Λ from the space of all possible experimental results to $\mathbb{R}$.

(2) Choose a significance threshold level $\alpha \in (0, 1)$. A popular choice is $\alpha = 5\%$, corresponding to a 95% confidence.

(3) Collect data (x) and evaluate $\Lambda(x)$.

(4) Calculate the p value (p) of $\Lambda(x)$. This p value is the probability of observing a value of Λ that is at least as extreme as $\Lambda(x)$ *if* the null hypothesis is true.

(5) Reject the null hypothesis if $p < \alpha$. Here, rejecting the null hypothesis means detecting context dependence.

Any procedure of this form ensures that the probability of falsely detecting context dependence is at most α. Within this constraint, it is desirable to choose a procedure—i.e., a statistic—with high *power* to detect context dependence if it is present. For general hypothesis testing, there is no universally optimal statistic except for the simplest problems [35], but the log-likelihood ratio (LLR) statistic is canonical and popular, and we find it to be convenient and powerful.

For data x, a statistical model parameterized by $\theta \in \mathcal{H}$ for some parameter space $\mathcal{H}$, and a null-hypothesis subspace $\mathcal{H}_0 \subset \mathcal{H}$, the LLR is defined as

$$\lambda \coloneqq -2\log[\mathcal{L}(\hat{\theta}_0)/\mathcal{L}(\hat{\theta})], \tag{3}$$

where $\mathcal{L}(\theta) = \Pr(\theta|x)$ is the likelihood function, $\hat{\theta}_0$ is the maximum likelihood estimate of θ over the null-hypothesis subspace $\mathcal{H}_0$, and $\hat{\theta}$ is the maximum likelihood estimate of θ over the full parameter space $\mathcal{H}$ [34–36]. For our problem, we have the following.

(1) $\mathcal{H}_0$: *the null hypothesis that* $\mathbf{p}_c = \mathbf{p}_0$ *for all* c *and some* $\mathbf{p}_0$.—The maximum likelihood estimate over the null hypothesis space is $\hat{\mathbf{p}}_0 = N^{-1}(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_M)$, with $\mathsf{x}_m = \sum_c \mathsf{x}_{c,m}$ counts obtained by aggregating over contexts, and $N = \sum_c N_c$.

(2) $\mathcal{H}$: *the full parameter space of independent* $\mathbf{p}_c$.—The maximum likelihood estimate in the full parameter space is $\hat{\mathbf{p}}_c = \mathbf{x}_c/N_c$.

Via basic multinomial statistics, the LLR is then

$$\lambda = -2\sum_{m=1}^{M}\left[\mathsf{x}_m \log\left(\frac{\mathsf{x}_m}{N}\right) - \sum_{c=1}^{C}\mathsf{x}_{c,m}\log\left(\frac{\mathsf{x}_{c,m}}{N_c}\right)\right]. \tag{4}$$

To compute p values, we appeal to Wilks' theorem [36]. It states that if the null hypothesis holds, as the number of samples $\to \infty$, then the LLR converges to a χ^2_k random variable, where $k = l - l_0$ and l (respectively, l_0) is the number of free parameters in the full (respectively, null) model [34–36]. Each probability vector contains $M-1$ free parameters (M probabilities summing to 1), so $l = C(M-1)$ and $l_0 = (M-1)$. If $N_c \gg 1$, then under the null hypothesis λ is approximately χ^2_k distributed, with

$$k = (C-1)(M-1). \tag{5}$$

The p value of an observed λ is therefore approximated by

$$p \approx 1 - F_k(\lambda), \tag{6}$$

where F_k is the χ^2_k cumulative distribution function. For prespecified α, we say that context dependence has been detected at significance α if $p < \alpha$. We call this simple primitive the *individual circuit test* (ICT), because it applies to data from a single circuit.

Here is a simple example of how the ICT can be used to detect context dependence. Consider a one-qubit circuit comprising the preparation of $|0\rangle$, application of $X_{\pi/2} = \exp(-i\pi\sigma_x/4)$, and measurement of σ_z. It is performed in two contexts: (1) while a neighbor qubit sits idle and (2) while the neighbor is driven in some fashion. Now, suppose the operations are perfect under context 1, but the driving in context 2 causes the $X_{\pi/2}$ gate to overrotate: $X_{\pi/2} \to \exp(-i1.1\pi\sigma_x/4)$. We choose a significance level of 5% and simulate 200 repetitions of the circuit in each context, observing 99 "0" outcomes in context 1 and 69 in context 2. Putting these data into Eqs. (4)–(6) with $C = 2$ and $M = 2$, we find that the p value is $p \approx 0.1\%$. This p value is easily significant at the 5% level ($p < 5\%$), so context dependence is detected in this simulated experiment. We also simulate a scenario where driving does *not* cause any change and this time obtain 108 "0" counts in context 1 and 107 in context 2. Calculated in the same way, the p value for these data is $p \approx 92\%$, so context independence is not rejected. If we repeat this simulation many times, in the latter case where there is no context dependence, we expect to erroneously detect context dependence in 5% of the trials.

B. Multicircuit data

Many experiments based on quantum circuits involve collecting data from multiple distinct circuits, as is the case for most QCVV techniques, including all RB protocols [5–21], GST [1–4], and other characterization methods [39,40]. We now extend the context dependence detection method presented above to the multicircuit scenario. Consider Q circuits indexed $q = 1, 2, \ldots, Q$, each with M possible outcomes, indexed $m = 1, 2, \ldots, M$ [41]. These circuits are all implemented in each of C contexts, again indexed by c for $c = 1, 2, \ldots, C$. Slightly generalizing the notation of Eq. (1), let

$$\mathbf{p}_{q,c} = (\mathsf{p}_{q,c,1}, \mathsf{p}_{q,c,2}, \ldots, \mathsf{p}_{q,c,M}) \tag{7}$$

denote the underlying probability distribution for circuit q in context c. As before, a particular circuit is context independent iff all $\mathbf{p}_{q,c} = \mathbf{p}_{q,0}$ for some circuit-dependent $\mathbf{p}_{q,0}$. All of the circuits are context independent if this holds for all circuits q.

Consider data generated by $N_{q,c}$ repeats of circuit q in context c. Let $\mathsf{x}_{q,c,m}$ denote count data for outcome m of circuit q in context c, with the full set of data denoted by

$$x = \{\mathbf{x}_{q,c} = (\mathsf{x}_{q,c,1}, \mathsf{x}_{q,c,2}, \ldots, \mathsf{x}_{q,c,M})\}. \tag{8}$$

There are many ways to test for context dependence with multicircuit data of this sort. Most obviously, we could apply the ICT defined above to the data from each circuit, to separately test for context dependence in each circuit. However, this means implementing Q statistical hypothesis tests. If the null hypothesis is true, and we naively implement Q independent hypothesis tests all at some fixed significance α, then we expect approximately αQ of the tests to falsely reject the null hypothesis just by random chance. In fact, the probability of falsely rejecting the null hypothesis in at least one test converges to 1 as Q increases.

To keep the probability of false detection in one or more tests—known as the *familywise error rate* (FWER) [35,42]—to at most α, it is necessary to adjust the significance of the individual tests. The simplest solution is the *generalized Bonferroni correction* [35,42]: For any tests implemented together, a FWER of at most α can be obtained by setting the "local" significance level of test i to $\alpha_i = \alpha w_i$ for any $w_i \geq 0$ satisfying $\sum_i w_i = 1$. Implementing all Q ICTs with each significance set to α/Q is therefore sufficient to maintain a global significance of α. However, the Bonferroni correction is unnecessarily conservative, so we use a strictly more powerful correction.

Because the λ_q are independent under the null hypothesis, where λ_q is the LLR for circuit q, we can implement the ICTs with a *Hochberg correction* [42–44]. In this setting, the Hochberg correction keeps the FWER to at most α using the following procedure.

(1) Order the Q p values from smallest to largest: $p_{(1)}, p_{(2)}, \ldots, p_{(Q)}$.

(2) Find the largest integer r such that $p_{(r)} \leq \alpha/(Q - r + 1)$, denoting this integer by $r_{\max}$.

(3) Reject the null hypothesis (context independence) for all circuits with p values smaller than

$$p_{\text{threshold}} = \alpha/(Q - r_{\max} + 1). \tag{9}$$

Hereafter, we always use this multiple-test correction procedure for the ICTs. Note that $p_{\text{threshold}}$ is not a true threshold for the statistical significance of a p value, in the sense that it depends on the data. We therefore refer to it instead as a "pseudothreshold." Sometimes, it is convenient to convert this to a pseudothreshold above which the LLR of a circuit is significant. Inverting Eq. (6), this pseudothreshold is given by

$$\lambda_{\text{threshold}} = F_k^{-1}(1 - p_{\text{threshold}}), \tag{10}$$

where k is the degrees of freedom per circuit, in Eq. (5), and F_k^{-1} is the inverse cumulative distribution function for the χ_k^2 distribution.

Controlling the FWER is not the only reasonable desideratum when implementing multiple hypothesis tests: A popular alternative is to control the *false discovery rate* (FDR), which is the expected ratio of the number of falsely-rejected null hypotheses to the total number of rejected null hypotheses. The FDR can be controlled to at most α in the ICTs using the Benjamini-Hochberg correction [45] (which has a similar form to the Hochberg correction). Whether controlling the FDR is preferable to controlling the FWER is subjective: Controlling the FDR increases test power at the cost of less certainty about the correctness of any specific positive ICT. We do not pursue this strategy herein (but our PYTHON implementation of these techniques [46] includes this alternative as an option).

The ICTs are often not the most sensitive for deciding whether there is context dependence in at least one circuit. In particular, there are tests that are more sensitive to context dependence that is distributed uniformly over all the circuits. A complementary test statistic, powerful for detecting uniformly distributed context dependence, is the *aggregate* LLR

$$\lambda_{\text{agg}} = \sum_{q=1}^{Q} \lambda_q, \tag{11}$$

where, again, λ_q is the LLR for circuit q. This is the LLR between the null hypothesis of context independence in *all* circuits and the full context dependence model. That is, it is the LLR between (1) the model defined by $\mathbf{p}_{q,c} = \mathbf{p}_{q,0}$ for some $\mathbf{p}_{q,0}$ and all q, and (2) the model in which all the $\mathbf{p}_{q,c}$ are independent. Therefore, when the null hypothesis holds, λ_{agg} approximately follows a $\chi^2_{k_{\text{agg}}}$ distribution with

$$k_{\text{agg}} = Q(C - 1)(M - 1). \tag{12}$$

For $k \gg 1$, the χ_k^2 distribution is approximately normal with mean k and variance $1/(2k)$. Therefore, in the common situations where $Q \gg 1$, it is convenient and intuitive to express the statistical significance of λ_{agg} by giving the number of standard deviations (N_σ) by which λ_{agg} exceeds it expected context-independent value. N_σ is given by

$$\mathcal{N}_\sigma = \frac{\lambda_{\text{agg}} - k_{\text{agg}}}{\sqrt{2k_{\text{agg}}}}. \tag{13}$$

In our experience, the p value of the aggregate LLR is often vanishingly small (see, e.g., Sec. IV), so $\mathcal{N}_\sigma$ provides an alternative measure of statistical significance that is on a more convenient scale. It is sometimes useful to have a threshold for α significance of the $\mathcal{N}_\sigma$, and this threshold is given by

$$\mathcal{N}_{\sigma,\text{threshold}} = \frac{F_{k_{\text{agg}}}^{-1}(1 - \alpha) - k_{\text{agg}}}{\sqrt{2k_{\text{agg}}}}. \tag{14}$$

When $Q \gg 1$, this threshold is essentially identical to the standard significance thresholds for standard deviations above the mean with a normal distribution.

Although the aggregate LLR test is often more sensitive, the ICTs are useful, because they indicate *which* circuits vary, which can constitute helpful diagnostic information, as demonstrated later. We can strike a balance between these tests by implementing the set of ICTs *and* the aggregate test, with significance levels adjusted appropriately. A reasonable strategy, which we adopt for the simulations and experiments in this paper, is the following. For a user-specified global significance α:

(1) Implement the aggregate test at significance level $\alpha/2$. If context dependence is detected, set $\beta = \alpha$; otherwise, set $\beta = \alpha/2$.

(2) Implement the ICTs using a Hochberg correction at a significance of β.

This multitest correction is based on the *closed test principle* (a generalization of the Bonferroni correction), and it controls the FWER to be at most α [47].

It is often useful to apply this entire procedure more than once while still maintaining a global significance of α. For example, if there are $C_{\rm all} > 2$ contexts in the data, we could choose to implement a pairwise comparison between multiple pairs of contexts (i.e., implementing this procedure more than once with $C = 2$), instead of—or as well as—implementing this procedure to jointly compare all contexts (i.e., one application of this procedure with $C = C_{\rm all}$). To maintain the global significance at α, we can perform a “top-level” Bonferroni correction, splitting α over each implementation of the procedure specified above. We use this strategy in Secs. IV and V, when applying these methods to simulated and experimental data.

C. Choosing the circuits

The context dependence detection methods that we propose in this section can be applied to data from almost any set of circuits. They can be bolted onto almost any device characterization protocol. However, if context dependence detection is a high priority, it is often useful to choose circuits that are sensitive to all the parameters that might vary with the context. GST circuits [1–4] are one reasonable choice, because they are informationally complete for tomography of gates, state preparations, and measurements (SPAM). If context dependence manifests as an observable dependence of gate or SPAM process matrices on the context, at least one GST circuit will be sensitive to it. If the effects of context dependence on gate behavior are small, then long-sequence GST (LSGST) can be used to amplify those effects [1], making them easier to detect. We use GST circuits in our examples below [using LSGST sequences for simulated data, and shorter linear-inversion GST (LGST) sequences [48] for experimental data, due to experimental constraints]. We do note, however, that in certain circumstances it may be easier or more desirable to run non-GST circuits, such as circuits prescribed by any of a number of randomized benchmarking protocols [5–21]. Additionally, it should be pointed out that, as we wish to determine if underlying probability distributions are identical or not, our empirical approximations of these distributions will become more accurate (and, hence, more sensitive to changes due to context dependence) as the number of times each circuit is repeated is increased, regardless of which circuits are used.

Using our tools on data from GST circuits does *not* require implementing the tomographic reconstructions of GST. Tomographic reconstructions using the data from each context are, nevertheless, clearly possible with GST data. This possibility naturally raises the question of what our tools add that could not be achieved as easily with tomography. Our tools have three distinct advantages over tomography, which highlight how they complement any tomographic data analysis. First, precise tomography require large amounts of data and many individual circuits, whereas detecting context dependence can often be achieved using few circuits and/or less data. Second, tomographic methods are based on fitting a model and become unreliable if this model does not accurately describe the system [25]. In contrast, these direct context dependence detection tools require no model of the underlying operations (the gates and SPAM).

III. QUANTIFYING CONTEXT DEPENDENCE

The detection methods presented in the previous section *test* whether or not there is statistically significant evidence of context dependence; when used rigorously, they report only “yes” or “no.” In general, the value of a test statistic will not necessarily quantify the “strength” of a detected effect. Neither the magnitude of the LLR for each circuit, nor the aggregate LLR, nor the associated p values, nor the aggregate $\mathcal{N}_\sigma$ directly quantify the strength of context dependence. Instead, they quantify our *confidence* that context dependence exists. If there is *any* context dependence in one or more circuits, then, as we take more data, both $\lambda_{\rm agg}$ and $\mathcal{N}_\sigma$ will increase without bound. A good quantitative measure of context dependence should describe the variations of an underlying gate or SPAM error rate, but measuring the variation in error rates is the domain of specific QCVV protocols (e.g., RB or GST). In the very general framework of this paper, the most we can do is to quantify the strength of each individual circuit’s context dependence, which is equivalent to estimating how much the circuit’s outcome probabilities change between contexts, and there are many ways to do this quantification.

A. Jensen-Shannon divergence

The simplest way to quantify context dependence is to rescale the per-circuit LLRs to

$$\mathrm{JSD}_q = \frac{\lambda_{\rm q}}{2N_q}, \tag{15}$$

where $N_q = \sum_c N_{q,c}$. As suggested by this notation, JSD_q provides an estimate of the Jensen-Shannon divergence (JSD) of the underlying probability distributions. For probability distributions P_c over M events, with $c = 1, 2, \ldots, C$, and some weightings π_c with $\sum_c \pi_c = 1$, the JSD is defined by [49]

$$\mathrm{JSD}_{\{\pi_c\}}(P_1, \ldots, P_C) = H\left(\sum_{c=1}^{C} \pi_c P_c\right) - \sum_{c=1}^{C} \pi_c H(P_c),$$

where $H(P)$ is the Shannon entropy of the probability distribution P given by

$$H(P) = -\sum_{m=1}^{M} P(m) \log P(m). \tag{16}$$

The JSD_q quantity defined in Eq. (15) is, in fact, the JSD (with a particular weighting) of the maximum likelihood estimates of the $\mathbf{p}_c$, so we call JSD_q the *observed JSD*. This equivalence can be shown directly by letting $P_c(m) \to \mathsf{x}_{c,m}/N_c$ and taking $\pi_c = N_c/N$ (where $N = \sum_c N_c$), in the definition of JSD.

The observed JSD is an estimate of the JSD of the underlying probability distributions for circuit q. Even if there is no context dependence, however, each JSD_q will almost always be nonzero due to ordinary finite-sample fluctuations. Thus, JSD_q is significantly different from zero only if it is greater than

$$\mathrm{JSD}_{\mathrm{threshold}} = \frac{\lambda_{\mathrm{threshold}}}{2N}, \tag{17}$$

where $\lambda_{\mathrm{threshold}}$ is the LLR pseudothreshold of Eq. (10). Implicit in this relation is the fact that λ_q and JSD_q are entirely equivalent test statistics.

B. Total variation distance

JSD quantifies statistical distinguishability between probability distributions and their average [49], so an estimate of the underlying JSD is a well-motivated measure of the context dependence of a circuit. However, there are other metrics with other meanings. One commonly used in quantum information is the total variation distance (TVD) [50]. The TVD between two distributions P_1 and P_2 over M events is

$$\mathrm{TVD}(P_1, P_2) = \frac{1}{2} \sum_{m=1}^{M} |P_1(m) - P_2(m)|. \tag{18}$$

The observed TVD for circuit q (TVD_q) is naturally defined by

$$\mathrm{TVD}_q = \frac{1}{2} \sum_{m=1}^{M} \left| \frac{\mathsf{x}_{1,m}}{N_1} - \frac{\mathsf{x}_{2,m}}{N_2} \right|. \tag{19}$$

Here, the contexts are indexed "1" and "2," because the TVD is defined only between two contexts, i.e., when $C = 2$.

Even if there is no context dependence, observed TVDs between two contexts are generally nonzero because of finite-sample fluctuations. It is often useful to correct for this result. Unlike the observed JSD, however, the observed TVD is not simply related to the LLR, so there is no simple pseudothreshold for TVD_q. Instead, we introduce the *statistically significant total variation distance* (SSTVD). If statistically significant variation is detected for circuit q using the ICTs, we report $\mathrm{SSTVD}_q = \mathrm{TVD}_q$ for that circuit; when no statistically significant context dependence is detected, the circuit has no SSTVD. That is,

$$\mathrm{SSTVD}_q = \begin{cases} \mathrm{TVD}_q & \text{if } \lambda_q > \lambda_{\mathrm{threshold}}, \\ \text{null} & \text{else}. \end{cases} \tag{20}$$

Note that we do not define SSTVD_q to be zero when $\lambda_q \leq \lambda_{\mathrm{threshold}}$. Just because we don't detect context dependence does *not* imply that no context dependence exists. Formally speaking, not rejecting a null hypothesis in a hypothesis test does not imply anything about whether that null hypothesis is true. For example, one or more λ_q could be just below the pseudothreshold at a global 5% significance and above the pseudothreshold at a global significance of 6%. Those circuits are, therefore, quite probably context dependent, meaning that a SSTVD_q of zero could be misleading.

When analyzing data from many circuits ($Q \gg 1$), it is often useful to summarize any observed context dependence with a single number. One such candidate is the maximum SSTVD over all circuits,

$$\max \mathrm{SSTVD} = \max_q [\mathrm{SSTVD}_q], \tag{21}$$

and we use this statistic in our examples later. The motivation for maxSSTVD is that it partially captures worst-case context dependence. maxSSTVD can be conveniently related to the diamond distance between gates, which is effectively a worst-case error rate; the diamond distance between channels $\mathcal{A}$ and $\mathcal{B}$ is a tight upper bound on the TVD (in circuit outcome probabilities, that is induced by replacing one instance of $\mathcal{A}$ with $\mathcal{B}$ in any circuit). (For further discussion of the diamond distance, see, e.g., Refs. [51–55].) Therefore, if SPAM operations are not context dependent, then the maximum *true* TVD over tested circuits, divided by the number of gates in the maximizing circuit, lower bounds the maximum diamond distance between corresponding gates in the two contexts. The maxSSTVD is an estimate of this maximal TVD and, therefore, roughly lower bounds the worst diamond distance between contexts' gate sets. (This link to the diamond distance suggests an interesting alternative to maxSSTVD:

$\max_q[\mathrm{SSTVD}_q/l(q)]$, where $l(q)$ is the length of circuit q.) It is also important to note that the value of max SSTVD is, in general, strongly dependent on the choice of circuits, even when divided by the circuit length, as the most context-dependent circuit might not be in the set of circuits chosen.

There are some subtleties to SSTVD, which can become important in slightly unusual circumstances. Perhaps the most significant of these is that the SSTVD of a circuit can *sometimes* significantly overestimate the true TVD of the circuit. For example, consider a situation whereby the TVD between contexts is the same and fairly small for all circuits and context dependence is detected in only some of the circuits (because the effect is small, so the chance that it is detected in any particular circuit is low). The circuits in which SSTVD is reported as non-null must have an observed TVD large enough so that the LLR test triggers, and the minimum such observed TVD could be significantly larger than the true TVD. If this is the case, any non-null SSTVD is a significant overestimate of the true TVD. Subtleties of this sort can be accounted for by looking at additional properties of the observed TVD distribution. However, this is not to suggest that looking at the full observed TVD distribution is always preferable in practice: The SSTVD is a convenient tool for highlighting the rough size of any detected context dependence without requiring subtle, case-specific analysis of a distribution.

IV. SIMULATED DRIFT DETECTION

In this section, we present a simulated example showing how to use the tools presented above to detect slow drift. This example uses data from GST circuits, but alternatives such as RB circuits could equally be used. We consider LSGST circuits [1] built from two gates: $\pi/2$ rotations around σ_x and σ_y. Each LSGST circuit begins with one of six short state-preparation sequences, followed by one of six short "germ" sequences repeated $O(K)$ times, and concludes with one of six short premeasurement sequences. These building blocks are chosen so that the collection of LSGST circuits is both informationally complete as well as amplificationally complete (they amplify sensitivity to all possible errors) [1,40]. Here, K ranges from 0 to 256 with logarithmic spacing, yielding 1405 unique quantum circuits. Below, the size of K is referred to as the "core" circuit length. The specific circuits used are given in the Appendix.

We simulate repeating these circuits $N = 100$ times in each of five consecutive time periods $t = 1, 2, \ldots, 5$ (the contexts). In addition to small time-*independent* unitary errors in the gates for the X and Y rotations [56], we simulate slow drift by adding overrotations of $(t-1) \times 10^{-3}$ rad in time periods t to both gates. We test for drift (context dependence between time periods) using a global significance level of $\alpha = 5\%$.

There are five contexts (the five time periods), so there are many ways to test for drift: We can implement the tests introduced in Sec. II B on all the data (jointly comparing the five contexts), and/or we can implement up to ten pairwise comparisons between pairs of different time periods (comparing pairs of contexts). We demonstrate all of these analyses, resulting in 11 comparisons between contexts in total. Therefore, to guarantee a global significance of 5%, we perform each comparison between contexts at a significance of $(5/11)\% \approx 0.45\%$ (this is a Bonferroni correction), with the aggregate LLR test and the ICTs performed for each comparison using the particular multitest correction procedure specified earlier. [That is, each aggregate LLR test is performed at $(5/22)\% \approx 0.23\%$ significance, leaving at least approximately 0.23% significance remaining for each collection of ICTs; the ICTs are then performed using the aforementioned Hochberg correction.] For the joint comparison of all five time periods, we find that the signed standard deviation of the aggregate LLR $\mathcal{N}_\sigma$, defined in Eq. (13), is $\mathcal{N}_\sigma \approx 21$; the threshold for drift detection is only $\mathcal{N}_\sigma \approx 2.9$ [as given by Eq. (14) with $\alpha \approx 0.23\%$]. Thus, we detect drift with extremely high confidence. The ICTs test also detects drift, finding 21 circuits to be significant.

To obtain more detailed, diagnostic information, we turn to the pairwise time period comparisons. These results are summarized in Fig. 2. The upper triangle in the upper plot in Fig. 2 shows $\mathcal{N}_\sigma$ for each pairwise comparison. For the longest time difference comparison, $\mathcal{N}_\sigma \approx 34$ (the threshold for drift detection is still $\mathcal{N}_\sigma \approx 2.9$). The lower triangle in the upper plot in Fig. 2 shows the number of circuits that are found to have statistically significant drift for each pairwise comparison. If this is zero *and* the $\mathcal{N}_\sigma$ is not statistically significant, then drift is not detected for that pairwise comparison; otherwise, it is. Therefore, none of the comparisons between neighboring time periods detect drift, but all other comparisons *do* detect drift. Drift is thus detected whenever the difference in the rotation angle between time periods is at least 2×10^{-3} rad. As expected, the statistical significance of the observed effect, as quantified by $\mathcal{N}_\sigma$, increases with the time delay. Note that, while no drift is detected between neighboring time periods, we know that drift is present (because we designed the model). This drift could have been made visible to our tools in either of two ways. First, we could have included longer sequences that would be more sensitive to small rotations. Alternatively, we could simply have collected more data.

Figure 2 also demonstrates that these tools allow for a rough diagnosis of the drift, without requiring computationally expensive exhaustive parameter estimation. The lower plot in Fig. 2 shows the distribution of the per-circuit observed JSDs, as defined in Eq. (15), versus the core circuit length (see above), for the longest delay period $t = 1$ versus $t = 5$. This comparison shows that the magnitude of the drift grows with the circuit length, implying that the gates are drifting, rather than the SPAM. Note that the only

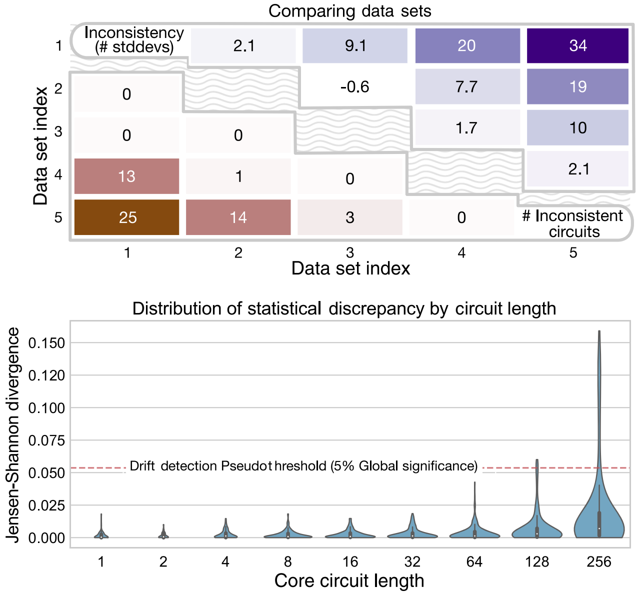

FIG. 2. An example using our techniques for drift detection on simulated data. Data are obtained by repeating the same 1405 circuits 100 times in each of five time periods. The circuits contain $\pi/2$ rotations around $\sigma_{x/y}$ and are informationally complete, meaning that they are collectively sensitive to drift in every aspect of gates and SPAM. Drift is modeled as time-dependent overrotations in both gates, by $(t-1)\times 10^{-3}$ rad in time period $t = 1, 2, \ldots, 5$. Upper plot, upper triangle: $\mathcal{N}_\sigma$ of total model violation for pairwise comparisons between the five pools. Upper plot, lower triangle: The number of circuits that are found to contain statistically significant drift. Lower plot: A violin plot of the estimated JSD for each circuit versus the core circuit length for the $t = 1$ to $t = 5$ time period comparison ("core" circuit length is defined in the main text). Any JSD above the pseudothreshold is significantly nonzero, at 5% global statistical significance, implying that drift is rigorously detected in the associated circuits. As discussed in the main text, by looking at which circuits have a high JSD, it is possible to infer the form of the errors.

circuits flagged by our tests as being context dependent at 5% global significance are those with an observed JSD above the pseudothreshold for statistical significance, given by Eq. (17) (there are 25 such circuits, as shown in the upper plot, and all are of length 128 or longer). Looking, however, at the trend in the observed JSD distribution for all circuits provides additional, if less rigorous, evidence of an increase in the underlying JSD with the circuit length (as more circuits have larger observed JSD at longer circuit length, even if most of those circuits' JSDs are still below the pseudothreshold) [57]. This evidence highlights the utility of further data analysis, after context dependence has been first detected with statistically rigorous hypothesis testing.

Looking at the specific details of the circuits, we observe that the largest observed JSDs are seen in circuits where the same gate is repeated sequentially many times. This observation strongly suggests that the gate rotation angles are drifting rather than the rotation axes (which those circuits would not amplify sensitivity to) or the stochastic error rates (changes in which would manifest in *all* longer sequences). This result is, of course, consistent with the simulated error model. Jupyter notebooks that contain this more detailed analysis, and which can be used to repeat and extend these simulations, are included as Supplemental Material [58].

V. EXPERIMENTAL DRIFT AND CROSSTALK DETECTION

To further demonstrate the practical utility of our tools, we applied them to detect and quantify drift and crosstalk in the publicly accessible ibmqx3 [38,59,60]. The ibmqx3, shown schematically in Fig. 3, is a 16-qubit superconducting device with connectivity on a 2×8 grid, resembling a ladder. We ran circuits over $\{\mathsf{I}, \mathsf{H}, \mathsf{S}\}$ gates on a single qubit (Q_{15}) to see whether

(I) the behavior of this qubit is affected by simultaneous CNOT gates applied to various "rungs" of the "ladder" or

(II) the behavior of this qubit drifts in time.

To do this experiment, we ran the circuits of LGST [48] over $\{\mathsf{I}, \mathsf{H}, \mathsf{S}\}$ on Q_{15} in multiple contexts. LGST is the simplest, least experimentally intensive form of GST, requiring only 40 unique circuits for these gates. The exact circuits are listed in the Appendix, and all the circuits are depth 7 or less. For each rung, we compare the output of

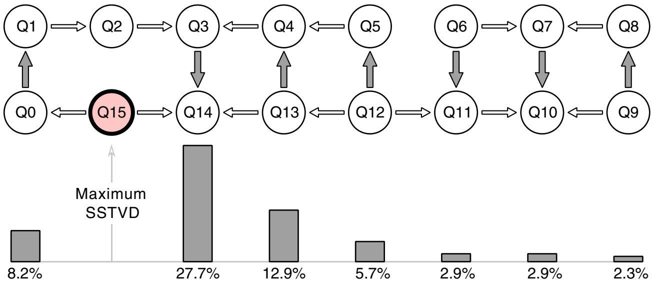

FIG. 3. Quantifying the effect of CNOT gates on the performance of qubit Q_{15} in ibmqx3 [38]. Top: A schematic of ibmqx3 with Q_{15} highlighted. Circles indicate qubits, and arrows denote CNOT gates, pointing from the control to the target. Bottom: The effect of driving each of the seven "ladder-rung" CNOT gates on short circuits run on qubit Q_{15}, as quantified by maxSSTVD, which is an empirical, total-variation-distance-based measure that we propose for estimating worst-case context dependence over circuits (see the main text). The maxSSTVD from driving each CNOT is plotted immediately below the corresponding rung in the schematic. The CNOT between qubits Q_{14} and Q_3 has a large effect on the behavior of circuits on Q_{15}, which corresponds to changing the outcome probabilities of a set of short circuits on Q_{15} by 27.7% in the worst case. The circuits run on Q_{15} are those of linear-inversion gate set tomography and are discussed in the main text.

LGST circuits on Q_{15} in the following time-ordered contexts.

(a) All other qubits idle.

(b) The CNOT on the rung is applied whenever a gate is applied to Q_{15}.

(c) All other qubits idle.

This experimental design is chosen to enable detection and isolation of both drift *and* crosstalk. If no context dependence is detected between (a) and (c), then we can safely rule out drift. Any context dependence between (a) and (b) may then be ascribed to crosstalk (modulo caveats discussed later). Access constraints prohibit running all the circuits for a rung in one submission. Therefore, for each rung, we submit the circuits for each context [(a)–(c)] in sequential batches. The delay between executed batches ranges from a few seconds to several minutes, depending on machine availability.

To implement the tests, we pick a global significance of 5%. To maintain this global significance level, a Bonferroni correction is used (following the prescription at the end of Sec. II B) to split this 5% evenly over the comparisons for the seven rungs and the (a) to (b) and (a) to (c) comparisons for each rung [we do not compare (b) to (c) so as to avoid additional local significance dilution]. This correction results in implementing each pairwise context comparison at a significance of $\frac{5}{14}\%$, noting that each pairwise comparison itself contains 40 per-circuit comparisons (the ICTs) and an aggregate comparison, as described earlier. As with the simulated data, the ICTs are performed with the Hochberg correction, as described in Sec. II B. (The resulting data, along with the full analysis, are provided in Supplemental Material [58].)

We detect no drift. That is, for all seven rungs, no change is detected between any (a) and corresponding (c) context. This result is interesting in its own right, but it is also critical for the crosstalk detection, which is because it implies that any variation between any (a) and (b) contexts is probably *not* due to random drift—and, thus, if differences are detected, that they are almost certainly due to the CNOT gate on the rung in question.

Our results comparing contexts (a) and (b) for each rung are summarized in Fig. 3, where we plot the maxSSTVD for each rung [see Eq. (21)]. In all cases, the application of CNOT gates on the other qubit pairs influences the behavior of Q_{15} to a statistically significant degree, as the maxSSTVD is nonzero [the SSTVD of a circuit is "null" if context dependence is not detected for that circuit; see Eq. (20)]. The observed maximum SSTVD broadly decreases with the connectivity graph distance between Q_{15} and the driven rung. Thus, closer CNOT gates generally affect Q_{15} more. For the CNOT between Q_3 and Q_{14}, one of the two closest rungs to Q_{15}, we observe a maxSSTVD of around 28%, corresponding to the gate sequence HSSSSH. For this circuit, out of 1024 measurement results, just two "1" outcomes are observed in context (a), while 286 "1" outcomes are observed in context (b). That is, this result suggests that applying the CNOT gate to this rung changes the outcome probabilities of this circuit on Q_{15} by about 28%.

The obvious cause of changes from contexts (a) to (b) is crosstalk, but there is an important caveat that needs to be addressed before we can draw this conclusion. The circuits on Q_{15} take longer when applying a CNOT to a rung [context (b)] than when implemented in isolation [context (a) or (c)]. This difference is because CNOT gates take substantially longer to implement than one-qubit gates on ibmqx3 [38], and in context (b) a single CNOT is applied in parallel with every gate acting on Q_{15}. Thus, a change in the output probabilities of Q_{15} from context (a) to (b) could be just due to the circuits taking longer, allowing for more decoherence to build up on Q_{15}.

This effect, however, is independent of the rung being tested, and this independence allows us to bound this effect. The maxSSTVDs between contexts (a) and (b) for the three furthest rungs are all approximately equal (see Fig. 3) and much lower than the maxSSTVDs for the other rungs. These maxSSTVDs provide a rough baseline for the maximal amount of the context dependence that can be attributed to this timing difference; any excess in the maxSSTVD above this level is almost certainly due to crosstalk.

To fully isolate the crosstalk caused by a CNOT from any change in circuit performance caused by an increased circuit duration, the time for each circuit layer should be fixed for all contexts, which could be more easily incorporated into experiments with lower-level access to a device. This desideratum is illustrative of the need to carefully account for all "nuisance contexts" that may be unintentionally or unavoidably changing with the context of interest. These nuisance contexts should be removed if possible or, as here, accounted for when not.

VI. DISCUSSION

To our knowledge, the tools we present and demonstrate herein are the first designed for detecting and characterizing generic context dependence in generic quantum circuits. However, one particular important example of context dependence is crosstalk, and there is already a widely used tool for characterizing crosstalk: simultaneous randomized benchmarking (SRB) [7,9]. For this reason, we now briefly discuss the relationship between our tools and SRB. In essence, SRB involves comparing a qubit's RB error rates in two contexts, corresponding to (1) leaving neighbor qubits idle and (2) driving them. This comparison then provides a quantification of crosstalk in terms of the increase in the RB error rate caused by driving neighboring qubits.

Our methods complement those of SRB: Our tools are not restricted to RB circuits, but, unlike SRB, they cannot directly provide a "crosstalk error rate" for the gates. Moreover, our methods cannot be applied directly to SRB data, because SRB uses independently sampled (and so almost certainly different) random sequences in each context. Our methods *can*, however, be used in concert with the SRB analysis if SRB is modified slightly, so that each random sequence appears in both the driven- and undriven-neighbor(s) contexts. With data from circuits of this sort, our tools complement the standard SRB analysis; they provide statistically rigorous crosstalk detection, something not directly addressed by the SRB analysis. Moreover, our tools allow for the testing of each *individual* random SRB sequence for sensitivity to driving, which can potentially help to identify the main sources of crosstalk (particularly if using varied-sampling-distribution RB methods such as those in Ref. [8]).

VII. CONCLUSIONS

Improving the performance of future quantum processors will require quantifying, understanding, and eventually mitigating a wide variety of context-dependent errors, such as crosstalk [7–9] and drift [23]. The techniques presented and demonstrated here are simple, general, and statistically rigorous ways to detect and quantify context-dependent errors, independent of their underlying physical causes. These methods are also computationally lightweight and can be applied to any collection of quantum circuits on any number of qubits. We therefore recommend that almost all device characterization protocols should be augmented with these tools. They can even be applied to archived data if any context-identifying information, such as time stamps, was kept. We expect that these techniques will contribute to the toolkit for calibrating and debugging next-generation qubits. For easy use, they have been integrated into (and documented in) the open-source PYGSTI software package [46].

ACKNOWLEDGMENTS

We thank Jay Gambetta, Lev Bishop, the IBM Quantum Experience team, and Erik Nielsen for technical support, and we thank Colm Ryan for alerting us to an overall scale error in Fig. 2(b) in the preprint version of this paper. The views expressed in the paper are strictly those of the authors and not of IBM. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract No. DE-NA0003525. This research was funded, in part, by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), and the U.S Department of Energy, Office of Advanced Scientific Computing Research Quantum Testbed Program. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

APPENDIX: CIRCUIT DETAILS

In this Appendix, we describe the sets of quantum circuits used in the simulations and experiments of the main text. The circuits are from two forms of GST [1–4]: LSGST [1] circuits are used for the simulations, while LGST [48] circuits are used for the experiments on ibmqx3. Below, we specify only the circuits used, not how this set of circuits is chosen. For more information on how to choose GST circuits, see Ref. [1] and the Jupyter notebooks accompanying this paper [58].

Following the notation of Ref. [1], the idle gate and gates corresponding to $\pi/2$ rotations around σ_x and σ_y are denoted by G_i, G_x, and G_y, respectively. The Hadamard and phase gates are denoted by G_h and G_s, respectively, where the phase gate is the unitary that maps $|x\rangle \rightarrow i^x|x\rangle$ for $x = 0, 1$. The null gate operation of "do nothing for no time" is denoted by "$\{\}$." Circuits are specified in operation order, *not* matrix multiplication order. For example, the sequence denoted G_hG_s means "perform a Hadamard gate, followed by a phase gate".

To succinctly list the circuits used in the simulations and experiments, it is necessary to first review the structure of GST circuits. Although not necessary, the GST circuits herein fix all state preparations to the $|0\rangle$ state and all measurements to be in the σ_z basis, so we specialize to that case. All GST circuits contain one of several short gate sequences at the beginning of the circuit, as well as another sequence at the end, which is to achieve tomographic completeness, by simulating informationally complete state preparations and measurements. These short sequences are referred to as *fiducials*. Given a gate set $\mathcal{G}$, a set of preparation fiducials $\mathcal{F}^{(p)}$, and a set of measurement fiducials $\mathcal{F}^{(m)}$, the collection of LGST circuits is the set of all circuits of the form

$$\begin{aligned}
&F, \quad \forall\ F \in \mathcal{F}^{(p)} \cup \mathcal{F}^{(m)},\\
&F_pF_m, \quad \forall\ F_p \in \mathcal{F}^{(p)}, \quad \forall\ F_m \in \mathcal{F}^{(m)},\\
&F_pGF_m, \quad \forall\ F_p \in \mathcal{F}^{(p)}, \quad \forall\ G \in \mathcal{G}, \quad \forall\ F_m \in \mathcal{F}^{(m)}.
\end{aligned}$$

Note that some circuits may appear more than once when iterating over all three forms of circuit and all possible combinations of gates, preparation fiducials, and measurement fiducials. (And, naturally, a circuit is added to the list of LGST circuits only once). From above, it follows that to define a set of LGST circuits it is necessary only to specify the sets $\mathcal{G}$, $\mathcal{F}^{(p)}$, and $\mathcal{F}^{(m)}$. For the experiment run on ibmqx3, we use the circuits of LGST with

$$\begin{aligned}\mathcal{G} &= \{G_i, G_h, G_s\},\\ \mathcal{F}^{(p)} &= \{\{\}, G_h, G_hG_s, G_hG_sG_s\},\\ \mathcal{F}^{(m)} &= \{\{\}, G_h, G_sG_h, G_hG_sG_h\}.\end{aligned}$$

In addition to the circuits of LGST, LSGST uses a further collection of sequences constructed from powers of a set of germs. Like the preparation and measurement fiducials, the germs are short sequences of gates from $\mathcal{G}$. Denote the germ set by $\mathbb{G}$, with the length of germ g denoted by $\ell(g)$. For LSGST, we also need to choose a maximum "germ power" $L_{\max} = 2^k$ for some positive integer k. LSGST consists of all the circuits of LGST along with all gate sequences of the form

$$F_p g^{\lfloor L/\ell(g)\rfloor} F_m, \quad \forall\ g \in \mathbb{G}, \quad \forall\ L \in \{1, 2, 4, \ldots, L_{\max}\},$$

where, as above, F_p and F_m run over all preparation and measurement fiducials, respectively. Again, these circuits may not all be unique or unique from the set of LGST circuits that they are combined with.

For the simulations presented in the main text to illustrate drift detection, we use LSGST circuits with $L_{\max} = 256$ and

$$\begin{aligned}\mathcal{G} &= \{G_x, G_y\},\\ \mathcal{F}^{(p)} = \mathcal{F}^{(m)} &= \{\{\}, G_x, G_y, G_x^2, G_x^3, G_y^3\},\\ \mathbb{G} &= \{G_x, G_y, G_xG_y, G_x^2G_y, G_xG_y^2, G_x^2G_yG_xG_y^2\},\end{aligned}$$

which results in 1405 circuits, as stated in the main text.

[1] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, *Demonstration of Qubit Operations below a Rigorous Fault Tolerance Threshold with Gate Set Tomography*, Nat. Commun. **8**, 14485 (2017).

[2] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, *Self-Consistent Quantum Process Tomography*, Phys. Rev. A **87**, 062119 (2013).

[3] D. Greenbaum, *Introduction to Quantum Gate Set Tomography*, arXiv:1509.02921.

[4] J. P. Dehollain *et al.*, *Optimization of a Solid-State Electron Spin Qubit Using Gate Set Tomography*, New J. Phys. **18**, 103018 (2016).

[5] 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).

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

[7] J. M. Gambetta *et al.*, *Characterization of Addressability by Simultaneous Randomized Benchmarking*, Phys. Rev. Lett. **109**, 240504 (2012).

[8] T. J. Proctor, A. Carignan-Dugas, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, *Direct Randomized Benchmarking for Multi-Qubit Devices*, arXiv:1807.07975.

[9] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, *Three Qubit Randomized Benchmarking*, arXiv:1712.06550.

[10] D. S. França and A.-L. Hashagen, *Approximate Randomized Benchmarking for Finite Groups*, J. Phys. A **51**, 395302 (2018).

[11] E. Magesan *et al.*, *Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking*, Phys. Rev. Lett. **109**, 080505 (2012).

[12] S. Sheldon, L. S. Bishop, E. Magesan, S. Filipp, J. M. Chow, and J. M. Gambetta, *Characterizing Errors on Qubit Operations via Iterative Randomized Benchmarking*, Phys. Rev. A **93**, 012301 (2016).

[13] T. Chasseur, D. M. Reich, C. P. Koch, and F. K. Wilhelm, *Hybrid Benchmarking of Arbitrary Quantum Gates*, Phys. Rev. A **95**, 062335 (2017).

[14] C. J. Wood and J. M. Gambetta, *Quantification and Characterization of Leakage Errors*, Phys. Rev. A **97**, 032306 (2018).

[15] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, *Characterizing Universal Gate Sets via Dihedral Benchmarking*, Phys. Rev. A **92**, 060302(R) (2015).

[16] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, *Real Randomized Benchmarking*, Quantum **2**, 85 (2018).

[17] W. G. Brown and B. Eastin, *Randomized Benchmarking with Restricted Gate Sets*, Phys. Rev. A **97**, 062323 (2018).

[18] J. Emerson, R. Alicki, and K. Życzkowski, *Scalable Noise Estimation with Random Unitary Operators*, J. Opt. B **7**, S347 (2005).

[19] J. Emerson, M. Silva, O. Moussa, C. Ryan, M. Laforest, J. Baugh, D. G. Cory, and R. Laflamme, *Symmetrized Characterization of Noisy Quantum Processes*, Science **317**, 1893 (2007).

[20] J. J. Wallman, M. Barnhill, and J. Emerson, *Robust Characterization of Loss Rates*, Phys. Rev. Lett. **115**, 060501 (2015).

[21] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, *Estimating the Coherence of Noise*, New J. Phys. **17**, 113020 (2015).

[22] A. Veitia, M. P. da Silva, R. Blume-Kohout, and S. J. van Enk, *Macroscopic Instructions vs Microscopic Operations*, arXiv:1708.08173.

[23] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak, *Nonexponential Fidelity Decay in Randomized Benchmarking with Low-Frequency Noise*, Phys. Rev. A **92**, 022326 (2015).

[24] M. D. Grace, J. M. Dominy, W. M. Witzel, and M. S. Carroll, *Optimized Pulses for the Control of Uncertain Qubits*, Phys. Rev. A **85**, 052313 (2012).

[25] S. J. van Enk and R. Blume-Kohout, *When Quantum Tomography Goes Wrong: Drift of Quantum Sources and Other Errors*, New J. Phys. **15**, 025024 (2013).

[26] C. Piltz, T. Sriarunothai, A. F. Varón, and C. Wunderlich, *A Trapped-Ion-Based Quantum Byte with 10-5 Next-Neighbour Cross-Talk*, Nat. Commun. **5**, 4679 (2014).

[27] C. Rigetti and M. Devoret, *Fully Microwave-Tunable Universal Gates in Superconducting Qubits with Linear*

Couplings and Fixed Transition Frequencies, Phys. Rev. B **81**, 134507 (2010).

[28] F. Altomare, K. Cicak, M. A. Sillanpää, M. S. Allman, A. J. Sirois, D. Li, J. I. Park, J. A. Strong, J. D. Teufel, J. D. Whittaker, and R. W. Simmonds, *Measurement Crosstalk between Two Phase Qubits Coupled by a Coplanar Waveguide*, Phys. Rev. B **82**, 094510 (2010).

[29] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, *Machine-Learning-Assisted Correction of Correlated Qubit Errors in a Topological Code*, Quantum **2**, 48 (2018).

[30] X.-M. Jin, Z.-H. Yi, B. Yang, F. Zhou, T. Yang, and C.-Z. Peng, *Experimental Quantum Error Detection*, Sci. Rep. **2**, 626 (2012).

[31] J. Florjanczyk and T. A. Brun, *In-Situ Adaptive Encoding for Asymmetric Quantum Error Correcting Codes*, arXiv:1612.05823.

[32] D. Greenbaum and Z. Dutton, *Modeling Coherent Errors in Quantum Error Correction*, Quantum Sci. Technol. **3**, 015007 (2018).

[33] D. Buterakos, R. E. Throckmorton, and S. D. Sarma, *Error Correction for Gate Operations in Systems of Exchange-Coupled Singlet-Triplet Qubits in Double Quantum Dots*, Phys. Rev. B **98**, 035406 (2018).

[34] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Science, New York, 2013).

[35] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses* (Springer Science, New York, 2006).

[36] S. S. Wilks, *The large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, Ann. Math. Stat. **9**, 60 (1938).

[37] A. Agresti, *Categorical Data Analysis*, 3rd ed. (Wiley, New York, 2012).

[38] 16-qubit backend: IBM Q team, "ibmqx3 backend specification," (2019) https://github.com/Qiskit/ibmq-device-information/tree/897c3d7a72a5deceafda66f9224e037f6c038ab0/backends/ibmqx3.

[39] S. Kimmel, G. H. Low, and T. J. Yoder, *Robust Calibration of a Universal Single-Qubit Gate Set via Robust Phase Estimation*, Phys. Rev. A **92**, 062315 (2015).

[40] K. Rudinger, S. Kimmel, D. Lobser, and P. Maunz, *Experimental Demonstration of Cheap and Accurate Phase Estimation*, Phys. Rev. Lett. **118**, 190502 (2017).

[41] The generalization to q-dependent M is avoided mostly only for notational simplicity.

[42] J. P. Shaffer, *Multiple Hypothesis Testing*, Annu. Rev. Psychol. **46**, 561 (1995).

[43] Y. Hochberg, *A Sharper Bonferroni Procedure for Multiple Tests of Significance*, Biometrika **75**, 800 (1988).

[44] More powerful corrections are also possible.

[45] Y. Benjamini and D. Yekutieli, *The Control of the False Discovery Rate in Multiple Testing under Dependency*, Ann. Stat. **29**, 1165 (2001).

[46] E. Nielsen, R. Blume-Kohout, L. Saldyt, J. Gross, T. Scholten, K. Rudinger, T. Proctor, and J. K. Gamble, pygsti version 0.9.7.2, 2019.

[47] F. Bretz, W. Maurer, W. Brannath, and M. Posch, *A Graphical Approach to Sequentially Rejective Multiple Test Procedures*, Stat. Med. **28**, 586 (2009).

[48] R. Blume-Kohout, J. K. Gamble, E. Nielsen, J. Mizrahi, J. D. Sterk, and P. Maunz, *Robust, Self-Consistent, Closed-Form Tomography of Quantum Logic Gates on a Trapped Ion Qubit*, arXiv:1310.4492.

[49] J. Lin, *Divergence Measures Based on the Shannon Entropy*, IEEE Trans. Inf. Theory **37**, 145 (1991).

[50] S. Verdú, *Total Variation Distance and the Distribution of Relative Information*, in *Proceedings of the 2014 Information Theory and Applications Workshop (ITA)* (2014), pp. 1–3, https://ieeexplore.ieee.org/document/6804281.

[51] J. Watrous, *Semidefinite Programs for Completely Bounded Norms*, Theory Comput. **5**, 217 (2009).

[52] Y. R. Sanders, J. J. Wallman, and B. C. Sanders, *Bounding Quantum Gate Error Rate Based on Reported Average Fidelity*, New J. Phys. **18**, 012002 (2015).

[53] J. Wallman, *Error Rates in Quantum Circuits*, arXiv:1511.00727v2.

[54] A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi, *Classical and Quantum Computation* (American Mathematical Society, Boston, 2002).

[55] D. Aharonov, A. Kitaev, and N. Nisan, *Quantum Circuits with Mixed States*, in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* (ACM, 1998), pp. 20–30, https://dl.acm.org/citation.cfm?id=276708.

[56] The random unitary errors induce, for the X and Y gates, respectively, diamond distances to ideal operations of 2.4×10^{-2} and 1.7×10^{-2}, with corresponding process infidelities of 5.8×10^{-4} and 2.8×10^{-4}. For further details, see Supplemental Material [58].

[57] Note that the correlation of the observed JSD with the circuit length could be turned into a test statistic and used in rigorous statistical hypothesis testing.

[58] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.9.021045 for datasets and Jupyter notebooks for reproducing the numerics presented in this paper.

[59] As of April 2019, ibmqx3 has been replaced by IBM Q 14 Melbourne, a 14-qubit device similar to ibmqx3.

[60] Qiskit, https://qisqkit.org.