Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit

Robin Blume-Kohout, John King Gamble, Erik Nielsen, Jonathan Mizrahi, Jonathan D. Sterk, and Peter Maunz
Sandia National Laboratories, Albuquerque, New Mexico 87185
(Dated: October 17, 2013)

We introduce and demonstrate experimentally: (1) a framework called "gate set tomography" (GST) for self-consistently characterizing an entire set of quantum logic gates on a black-box quantum device; (2) an explicit closed-form protocol for linear-inversion gate set tomography (LGST), whose reliability is independent of pathologies such as local maxima of the likelihood; and (3) a simple protocol for objectively scoring the accuracy of a tomographic estimate without reference to target gates, based on how well it predicts a set of testing experiments. We use gate set tomography to characterize a set of Clifford-generating gates on a single trapped-ion qubit, and compare the performance of (i) standard process tomography; (ii) linear gate set tomography; and (iii) maximum likelihood gate set tomography.

Quantum information processing (QIP) relies upon precise, repeatable quantum logic operations. Experiments in multiple QIP technologies [1–5] have implemented quantum logic gates with sufficient precision to reveal weaknesses in the *quantum tomography* protocols used to characterize those gates. Conventional tomographic methods assume and rely upon a precalibrated reference frame, comprising (1) the measurements performed on unknown states, and (2) for quantum process tomography, the test states that are prepared and fed into the process (gate) to be characterized. Standard process tomography on a gate G proceeds by repeating a series of experiments in which state ρ_j is prepared and observable (a.k.a. *POVM effect*) E_k is observed, using the statistics of each such experiment to estimate the corresponding probability

$$p_{k|j} = \mathrm{Tr}[E_k G[\rho_j]]$$

(given by Born's rule), and finally reconstructing G from many such probabilities.

But, in most QIP technologies, the various test states (ρ_j) and measurement outcomes (E_k) are *not* known exactly. Instead, they are implemented using the very same gates that process tomography is supposed to characterize. The quantum device is effectively a black box, accessible only via classical control and classical outcomes of quantum measurements, and in this scenario standard tomography can be dangerously self-referential. If we do process tomography on gate G under the common assumption that the test states and measurement outcomes are both eigenstates of the Pauli $\sigma_x, \sigma_y, \sigma_z$ operators, then the accuracy of the estimate $\hat{G}$ will be limited by the error in this assumption.

This is now a critical experimental issue. In platforms including (but not limited to) superconducting flux qubits [1], trapped ions [5], and solid-state qubits, quantum logic gates are being implemented so precisely that systematic errors in tomography (due to miscalibrated reference frames) are glaringly obvious. Fixes have been proposed [1, 2, 6, 7], but none yet provide a general, comprehensive, reliable scheme for gate characterization.

FIG. 1: **The GST model of a quantum device.** Gate set tomography treats the quantum system of interest as a black box, with strictly limited access. This is a fairly good model for many qubit technologies, especially those based on solid state and/or cryogenic technologies. We do not have direct access to the Hilbert space or any aspect of it. Instead, the device is controlled via *buttons* that implement various gates (including a preparation gate and a measurement that causes one of two indicator lights to illuminate). Prior information about the gates' function may be available, and can be used, but should not be *relied* upon.

In this article, we present *gate set tomography* (GST), a complete scheme for reliably and accurately characterizing an entire set of quantum gates. In particular we introduce the first *linear-inversion* protocol for self-consistent process tomography, linear gate set tomography (LGST). LGST is a closed-form estimation protocol (inspired in part by [8–10]) that cannot – unlike pure maximum-likelihood (ML) algorithms – run afoul of local maxima in a likelihood function that is gener-

ally ill-behaved. While the price of LGST's reliability is decreased accuracy compared with ML, it is easy to recover accuracy using a hybrid scheme in which the LGST estimate is used as the starting point for local ML estimation. We demonstrate (L)GST experimentally by characterizing a complete set of gates for a trapped-ion qubit. To demonstrate its performance, we introduce a novel quantitative scoring protocol that evaluates how well a tomographic estimate predicts independent "test" experiments.

I. BACKGROUND

We begin with a brief review of standard tomography, and the mathematical conventions used in both standard and gate set tomography.

A. Mathematical conventions

A quantum information processing device is described using a Hilbert space $\mathcal{H}$ of finite dimension d (any system with $d = 2$ is a *qubit*). Its state at any time is described by a $d \times d$ *density matrix* ρ that is positive semidefinite ($\rho \geq 0$) and has unit trace ($\mathrm{Tr}\rho = 1$). Each possible measurement on the system is represented by a *positive, operator-valued measure* (POVM), comprising a set $\mathcal{M} = \{E_k\}$ of $d \times d$ matrices that are positive semidefinite ($E_k \geq 0$) and sum to the identity matrix ($\sum_k E_k = \mathbb{1}$). This representation gains its meaning from Born's rule, which states that when measurement $\mathcal{M}$ is performed on a system in state ρ, outcome "k" will be observed with probability

$$Pr(k|\rho) = \mathrm{Tr}[E_k\rho].$$

A system's reversible evolution is described by some $d \times d$ unitary operator U, and the state evolves as $\rho_{t_2} = U\rho_{t_1}U^\dagger$. In practice, a system's dynamics (e.g., when a logic gate is applied in the laboratory) will be at least slightly irreversible, and must be represented by a *completely positive, trace-preserving linear map on density matrices* (CPTP map) that can be written in Kraus form:

$$G[\rho] = \sum_i K_i \rho K_i^\dagger,$$

where the $\{K_i\}$ are matrices satisfying $\sum_i K_i K_i^\dagger = \mathbb{1}$.

In tomography, it is more useful to represent quantum processes using the *Hilbert-Schmidt space* of matrices on $\mathcal{H}$, denoted $B(\mathcal{H})$, in which any $d \times d$ matrix X is a column vector $|X\rangle\!\rangle$ or row vector $\langle\!\langle X|$. In this representation, we can write Born's Rule as

$$Pr(k|\rho) = \langle\!\langle E_k|\rho\rangle\!\rangle \tag{1}$$

using the Hilbert-Schmidt inner product $\langle\!\langle X|Y\rangle\!\rangle \equiv \mathrm{Tr}[X^\dagger Y]$. Since quantum processes are always linear maps on density matrices, they can always be represented as $d^2 \times d^2$ matrices (a.k.a. superoperators) on Hilbert-Schmidt space. In this representation, if a process G is applied to a state ρ and then a measurement $\mathcal{M} = \{E_k\}$ is performed, then the probability of observing outcome "k" is simply

$$Pr(k|\rho, G) = \langle\!\langle E_k|\, G\, |\rho\rangle\!\rangle.$$

B. Standard tomography

In this framework, quantum state tomography [11–13] is a simple linear algebraic inversion. Given an unknown state $|\rho\rangle\!\rangle$, we characterize it by performing a set of measurements that are *informationally complete* – i.e., their outcomes $\{E_k\}$ collectively span $B(\mathcal{H})$. We repeat the measurement(s) N times, count the observations of outcome "k" (n_k), estimate its probability as

$$Pr(k|\rho) \approx \frac{n_k}{N} \equiv \hat{p}_k, \tag{2}$$

and use linear algebra to invert the set of equations

$$\langle\!\langle E_k|\rho\rangle\!\rangle = \hat{p}_k. \tag{3}$$

Quantum process tomography [13, 14] is very similar, but in addition to an informationally complete set of measurement outcomes, we must also prepare a set of test states ρ_j that span $B(\mathcal{H})$, and apply the unknown process G to them. From the count statistics of these repeated experiments, we estimate

$$Pr(k|G, \rho_j) \approx \frac{n_{j,k}}{N} \equiv \hat{p}_{j,k}, \tag{4}$$

and use linear algebra to invert the set of equations

$$\langle\!\langle E_k|\, G\, |\rho_j\rangle\!\rangle = \hat{p}_{j,k}. \tag{5}$$

These techniques define *linear-inversion tomography*. While they have been largely supplanted by more sophisticated statistical methods [15–17] that provide better accuracy for finite N, linear-inversion techniques remain useful in the limit $N \to \infty$. More importantly, their existence guarantees the existence of efficient, reasonably accurate protocols for standard tomography. They can serve as a first stage in more accurate protocols [18], or enable existence proofs for more sophisticated protocols [19, 20].

These protocols, however, assume that the $\{\rho_j\}$ and $\{E_k\}$ are known – and use them to define an absolute reference frame for the quantum system's state space. This is a theorist's fiction; in every quantum technology except (arguably) linear optics, no such reference frame is given. The assortment of test states and measurements required for process tomography are obtained by applying the same dynamical gates that process tomography is supposed to characterize to a *single* imperfectly known fiducial state ρ and a single imperfectly known fiducial measurement $\mathcal{M}$.

II. GATE SET TOMOGRAPHY

Gate set tomography (GST) is based on a simple insight: playing around with a quantum device should be sufficient to reveal all the properties needed to predict its future behavior. If this is true, then every unjustified assumption should also be unnecessary. We need not – and should not – assert that certain operations prepare $|0\rangle$ states, measure the σ_z basis, etc. If they do operate this way, then the data will reveal it.

To enforce this intellectual discipline, we model the quantum device as a black box (see Fig. 1, and also Ref. [21], which pioneered the idea that black box qubits should be fully characterizable). Our interaction with the black box is strictly classical and limited to pushing a small number of "buttons" (generally implemented in experiments by electromagnetic control pulses):

- One button, marked "ρ", initializes the system.
- Another button, marked "$\mathcal{M}$", performs a 2-outcome measurement – it is accompanied by 2 lights, exactly one of which lights up to indicate the outcome.
- Finally, a set of K buttons labeled $G_1 \ldots G_K$ perform quantum operations (logic gates) on the system.

All of these buttons' effects are unknown, and have to be deduced from the data. No other controls exist. In this article, we will make a number of simplifying assumptions, all of which can be relaxed (at some cost – which will be discussed further in [22]) to make GST more robust.

- The Hilbert space dimension $d = \dim(\mathcal{H})$ is known.
- The effect of the initialization button really is to reprepare the system (repeatably) in a state ρ.
- The measurement button $\mathcal{M}$ can be represented by a 2-outcome POVM $\{E, \mathbb{1} - E\}$.
- Control is Markovian: Each button can be represented by a completely positive, trace-preserving (CPTP) map on $B(\mathcal{H})$.

A *gate set*, then, is a complete description of a black box. In Hilbert-Schmidt space notation, it is

$$\mathcal{G} = \{|\rho\rangle\rangle, \langle\langle E|, \{G_k\}\}.$$

The goal of GST is to identify $\mathcal{G}$ from the results of experiments on the black box system.

A. Experiments, data, and inference in GST

GST, like all tomography protocols, comprises (1) obtaining data, and (2) analyzing the data to get an estimate. Data are gathered from a discrete set of M *experiments*, each of which is repeated many (N) times to get statistics. Experiments have a simple form:

1. Push the "ρ" button to initialize the system.
2. Apply a sequence $s = \{G_{s_1}, G_{s_2}, G_{s_3}, \ldots, G_{s_L}\}$ of L gates.
3. Push the "$\mathcal{M}$" button and record the outcome.

Experiments are described and indexed by the sequence s, and the data comprise the observed counts $\{n_s\}$ (for each of the M values of s that were performed). Note that sequences of gates are equivalent to quantum circuits, except that in circuit design it is usually assumed that certain gates commute (i.e., gates on different qubits) and can be performed in parallel. Since this may be violated in experimental hardware, we do not assume it in GST – if two gates *do* commute, it will be apparent in the data. Since sequences correspond to circuits, we can see GST as predicting the statistics of *arbitrary* circuits by studying the behavior of a *specific* (and limited) set of circuits.

Each experiment s has two outcomes, and is thus associated with a single probability

$$p_s = Pr(E|\rho, s) = \langle\langle E| G_{s_L} \circ G_{s_{L-1}} \circ \ldots \circ G_{s_2} \circ G_{s_1} |\rho\rangle\rangle.$$

That experiment's observed counts (n_s) provide information about p_s, which is a single parameter of the gate set $\mathcal{G}$. A simple if crude inference procedure is to estimate

$$\hat{p}_s = \frac{n_s}{N}, \tag{6}$$

and thus to nail down the parameters of $\mathcal{G}$ one by one. If $\dim(\mathcal{H}) = d$, then ρ requires $d^2 - 1$ parameters, E requires d^2, and each of the K gates G_k requires $d^2(d^2 - 1)$, suggesting that

$$M \approx Kd^4 - (K-2)d^2 - 1$$

distinct experiments should be necessary and sufficient to identify $\mathcal{G}$.

B. The gauge

Not every parameter in $\mathcal{G} = \{\rho, E, \{G_k\}\}$ can be estimated, however. Given a gate set $\mathcal{G}$, let M be some invertible $d^2 \times d^2$ superoperator, and suppose that we construct a different gate set $\mathcal{G}'$ given by

$$\begin{aligned} |\rho'\rangle\rangle &= M |\rho\rangle\rangle \\ \langle\langle E'| &= \langle\langle E| M^{-1} \\ G'_k &= M G_k M^{-1}. \end{aligned} \tag{7}$$

Every observable probability $p_s = \langle\langle E| G_{s_L} \circ \ldots \circ G_{s_1} |\rho\rangle\rangle$ is identical for $\mathcal{G}$ and $\mathcal{G}'$. So the action of M is a *gauge transformation* (see also [1]), and $\mathcal{G}'$ and $\mathcal{G}$ are equivalent representations of the same physical gate set. The gauge group is $SL(d^2)$, since M must be invertible, and M and αM act identically for any scalar α.

This gauge freedom means that the standard representation of a gate set as $\{\rho, E, \{G_k\}\}$ (which coincides with the way that operations and states are conventionally represented in quantum information) contains approximately $d^2 - 1$ redundant and unobservable parameters. Instead of describing the system's observable dynamics, they define only the convenient but arbitrary reference frame (akin to the conventional $\hat{x}, \hat{y}, \hat{z}$ axes in space) in which a given experimentalist or theorist has chosen to express those dynamics. For example,

$$\mathcal{G}_0 = \{\rho = E = |0\rangle\langle 0|, G_1 = e^{i\sigma_z \pi/4}, G_2 = e^{i\sigma_x \pi/4}\}$$

is indistiguishable from

$$\hat{\mathcal{G}} = \{\rho = E = |+\rangle\langle +|, G_1 = e^{i\sigma_x \pi/4}, G_2 = e^{i\sigma_y \pi/4}\}.$$

The only difference is what we have chosen to call the computational basis.

We do not yet know any satisfying gauge-invariant "normal form" for gate sets (although some of the gauge parameters can be fixed in obvious ways, e.g. by defining ρ to be diagonal in the computational basis), nor a well-motivated gauge-invariant measure of fidelity between gate sets. The sets $\mathcal{G}_0$ and $\hat{\mathcal{G}}$ shown above appear quite different, and the gate-by-gate fidelity between them would be very low by any measure. Yet they are in fact indistinguishable. If an experimentalist set out to implement $\mathcal{G}_0$, it would be quite unfair if a tomographer reported that (1) in fact $\hat{\mathcal{G}}$ was being implemented, and therefore (2) the fidelity of implementation is quite low! Ultimately, we aspire to a gauge-invariant theory, or at least a canonical way of fixing the gauge. In its absence, we compare a tomographic estimate $\hat{\mathcal{G}}$ to a target $\mathcal{G}_0$ by optimizing the gauge numerically.

To compare $\hat{\mathcal{G}}$ with $\mathcal{G}_0$, we search for the gauge transformation $M \in SL(d^2)$ that makes $\hat{\mathcal{G}}$ as similar as possible to $\mathcal{G}_0$ (e.g., as measured by $\sum_k ||G_k - \hat{G}_k||_2^2$). Obviously, scientific integrity suggests that the intended target should play no role in the estimation of physically observable quantities. So first, we perform tomography and obtain an estimate $\hat{\mathcal{G}}$ without considering the gauge (or the target). Only then do we vary the gauge in which $\hat{\mathcal{G}}$ is described (which has no effect on anything observable) to minimize the discrepancy between $\mathcal{G}_0$ and $\hat{\mathcal{G}}$.

C. Positivity

Positivity is a highly desirable property of any theory; it means that no matter what weird objects appear internal to the theory, every observable probability is always in the range $[0, 1]$. In the conventional representation of quantum operations, positivity demands that:

- ρ is a positive semidefinite operator with trace 1,
- E and $\mathbb{1} - E$ are positive semidefinite,
- Each G_k is a CPTP map [i.e., $(G_k \otimes \mathbb{1})[\rho] \geq 0 \; \forall \; \rho \geq 0$].

These conditions are always sufficient for positivity, but not strictly necessary in the black box model. In the black box model, we cannot prepare arbitrary states (so E need not be strictly positive), nor perform arbitrary measurements (so ρ need not be strictly positive), nor inject systems that are entangled with external ancillae (so complete positivity is something of a red herring). But since we reasonably anticipate that quantum mechanics is the same inside the black box as outside, it is reasonable to demand that our estimate satisfy the conventional positivity conditions anyway.

However, gauge transformations do not respect the conventional positivity constraints. Gauge transformation of a gate set in which every gate is CPTP can easily yield a gate set in which several (if not all) of the gates violate complete positivity. It is natural to define a CPTP gate set as one that is gauge-equivalent to a set of CPTP gates, but we have no closed-form test for this property. An alternative is to demand that each gate G_k be explicitly CP, but this constraint truncates the gauge freedom. For an extremal gate set – where ρ and E are rank-1 projectors, and each G_k is a unitary – complete positivity simply reduces the gauge group to $SU(d)$. Every gauge transformation that does not lie in this subgroup would produce a new gate set that violated positivity constraints.

For noisy gate sets, in which each gate lies in the interior of the set of CPTP maps, requiring positivity has more complicated consequences for the gauge. Any sufficiently small $SL(d^2)$ transformation will preserve positivity. But if a gauge transformation outside of the $SU(d)$ subgroup is iterated enough times, then it will eventually violate positivity. Gauge transformations do not form a group. There may exist pairs of CPTP gate sets that are gauge-equivalent, yet are not connected by a continuous path of gauge transformations.

These complications are severe enough that in this work, we do *not* impose complete positivity, much as as early work on linear-inversion state tomography did not impose positive semidefiniteness on the density matrix $\hat{\rho}$. Instead, we allow the estimated gates to be arbitrary matrices, and rely on consistency with experimental data to ensure that the gate sets predict positive probabilities for future experiments. Careful implementation of positivity constraints is a clear and pressing subject for future work.

D. Practical inference of the gate set

Estimating $\hat{p}_s = \frac{n_s}{N}$ and reconstructing $\mathcal{G}$ from these estimated probabilities is impractical. For one thing, the linear inversion estimate of p_s is imperfect (n_s is rarely equal to $p_s N$). A more elegant and robust approach is

to define a *likelihood function* over gate sets,

$$\mathcal{L}(\mathcal{G}) = \prod_s p_s(\mathcal{G})^{n_s}(1-p_s(\mathcal{G}))^{N-n_s}, \tag{8}$$

and construct an estimate from it. A simple and popular (albeit still suboptimal; see discussion in [16]) technique is maximum likelihood estimation, which reports

$$\hat{\mathcal{G}}_{ML} = \text{argmax}[\mathcal{L}(\mathcal{G})].$$

But here, GST (and other techniques for self-consistent gate estimation) diverge from standard state and process tomography. In standard tomography the likelihood function is log-convex [16], and therefore has a unique local maximum that can be found via a variety of numerical techniques. But in GST, the observable probabilities $p_s(\mathcal{G})$ are not linear functions of the parameters of $\mathcal{G}$. They are polynomials of degree L, because each gate can appear up to L times in an experiment (e.g., $p = \langle\langle E| G_1^L |\rho\rangle\rangle$). This makes the crude approach of estimating the probabilities $\{\hat{p}_s\}$ and then solving for $\mathcal{G}$ almost impossible, but (more worryingly), it also means that the GST likelihood function will not generally be log-convex or have a unique local maximum.

Prior approaches made use of the assumption that the target gates $\mathcal{G}_0$ are a good prior estimate of $\mathcal{G}$, e.g. by maximizing a series expansion of $\mathcal{L}(\mathcal{G})$ in the neighborhood of $\mathcal{G}_0$ [1]. This may work in many cases, but it depends critically on the accuracy of the prior knowledge – and could lead to worryingly circular estimates (i.e., MLE may find a local maximum near the prior estimate, even if the prior estimate is wildly wrong and the true global maximum is far away). In the next section, we solve this problem with a robust, closed-form estimator that can be used directly, or as a reliable "pretty close" starting point for MLE.

III. LINEAR GATE SET TOMOGRAPHY

Linear inversion state tomography, the oldest and simplest form of tomography [11, 14], is based on the notion that we should assign an estimate $\hat{\rho}$ that predicts probabilities equal to observed frequencies:

$$\text{Tr}(\hat{\rho}E) = Pr(E|\hat{\rho}) = \frac{n_E}{N}. \tag{9}$$

When such a $\hat{\rho}$ (1) exists and (2) is physically valid, it will also maximize the likelihood. So linear inversion and MLE coincide in such cases. When the data are overcomplete, the set of equations implied by Eq. 9 are overconstrained. Linear inversion is still possible using least-squares inversion, which minimizes a weighted sum of residuals between probabilities and observed frequencies,

$$\text{Err}(\rho) = \sum_k w_k \left(\text{Tr}(\hat{\rho}E_k) - \frac{n_E}{N}\right)^2. \tag{10}$$

However, the (often neglected) weights w_k are actually rather important, since they determine which of the conflicting observations will dominate. To figure out what these weights should be, we are generally forced to turn to MLE anyway, and the best weighted least-squares fit is simply the argmax of a Gaussian approximation to the likelihood function. For these reasons, linear inversion has been largely replaced by MLE.

Linear inversion nonetheless remains not only a powerful conceptual tool, but also the only closed-form tomographic protocol. It proves that pretty good state tomography can in fact be done efficiently. This has never been in doubt – but gate set tomography is a different kettle of fish. It is *not* obviously feasible, for the likelihood function is not necessarily unimodal because event probabilities depend nonlinearly on the gate-set parameters.

We remedy this problem here by presenting a simple method for linear inversion gate set tomography (LGST), and a closed-form expression for the estimate. Our approach makes implicit use of a Gram matrix technique similar to that used by Cyril Stark in [8–10]. We do not propose raw LGST as a final estimator – it is clearly suboptimal in accuracy – but as a critical part of a toolchain. It (1) proves in principle that efficient closed-form GST is possible, and (2) provides in practice a good starting point for gradient-based likelihood maximization.

A. Derivation of the LGST algorithm

Let $\{F_k\}_{k=1\ldots d^2}$ be a set of quantum operations, each implemented by a short "fiducial" gate string:

$$F_k = G_{f_k(L)} \circ G_{f_k(L-1)} \circ \ldots G_{f_k(1)} \tag{11}$$

Now, using the fixed and unknown state ρ and effect E, let us define

$$\begin{aligned} |\rho_k\rangle\rangle &= F_k |\rho\rangle\rangle \\ \langle\langle E_k| &= \langle\langle E| F_k, \end{aligned} \tag{12}$$

and, in terms of them, the (unknown) matrices

$$\begin{aligned} A &= \sum_j |j\rangle\rangle\langle\langle E_j| \\ B &= \sum_k |\rho_k\rangle\rangle\langle\langle k|. \end{aligned} \tag{13}$$

Next, for any gate X, define

$$\begin{aligned} \tilde{X}_{jk} &= \langle\langle E_j| X |\rho_k\rangle\rangle \\ &= \langle\langle E| F_j X F_k |\rho\rangle\rangle \end{aligned} \tag{14}$$

Since $\tilde{X}_{jk}$ is the probability of an experimentally observable event corresponding to the sequence $F_j X F_k$, we can "measure" $\tilde{X}_{jk}$ to whatever accuracy we desire, and construct the matrix

$$\tilde{X} = \sum_{j,k} |j\rangle\rangle\langle\langle k| \tilde{X}_{jk}. \tag{15}$$

Now, although we do not know the matrices A and B, we observe that

$$\tilde{X} = AXB, \tag{16}$$

and in particular

$$\tilde{\mathbb{1}} = AB. \tag{17}$$

The final, critical observation is that if $\tilde{\mathbb{1}}$ is invertible, then $\tilde{\mathbb{1}}^{-1} = B^{-1}A^{-1}$ and

$$\tilde{\mathbb{1}}^{-1}\tilde{X} = B^{-1}A^{-1}AXB = B^{-1}XB. \tag{18}$$

So, for each gate G_i, we define

$$\hat{G}_i = \tilde{\mathbb{1}}^{-1}\tilde{G}_i. \tag{19}$$

This is (ignoring statistical fluctuations) a perfectly good estimate, since $\hat{G} = \{B^{-1}G_iB\}$ is gauge-equivalent to $G = \{G_i\}$. To estimate $|\rho\rangle\!\rangle$ and $\langle\!\langle E|$, we define the (element-wise identical) vectors

$$|\tilde{\rho}\rangle\!\rangle = A\,|\rho\rangle\!\rangle = \sum_j |j\rangle\!\rangle\,\langle\!\langle E|\,F_j\,|\rho\rangle\!\rangle \tag{20}$$

$$\left\langle\!\langle\tilde{E}\right| = \langle\!\langle E|\,B = \sum_k \langle\!\langle E|\,F_k\,|\rho\rangle\!\rangle\,\langle\!\langle k|, \tag{21}$$

and observe that that linear-inversion estimates in the same gauge as the $\hat{G}_k$ estimates can be obtained as

$$|\hat{\rho}\rangle\!\rangle = \tilde{\mathbb{1}}^{-1}\,|\tilde{\rho}\rangle\!\rangle = B^{-1}\,|\rho\rangle\!\rangle \tag{22}$$

$$\left\langle\!\langle\hat{E}\right| = \left\langle\!\langle\tilde{E}\right| = \langle\!\langle E|\,B. \tag{23}$$

B. How to implement LGST

The procedure for LGST is therefore to repeatedly perform each of the experiments

$$\langle\!\langle E|\,F_j \circ G_i \circ F_k\,|\rho\rangle\!\rangle,$$
$$\langle\!\langle E|\,F_j \circ F_k\,|\rho\rangle\!\rangle,$$
$$\langle\!\langle E|\,F_j\,|\rho\rangle\!\rangle,$$

gather statistics to estimate their probabilities, arrange those probabilities into matrices as

$$\tilde{\mathbb{1}} = \sum_{j,k} \langle\!\langle E|\,F_jF_k\,|\rho\rangle\!\rangle\,|j\rangle\!\rangle\langle\!\langle k|\,, \tag{24}$$

$$\tilde{G}_i = \sum_{j,k} \langle\!\langle E|\,F_jG_iF_k\,|\rho\rangle\!\rangle\,|j\rangle\!\rangle\langle\!\langle k|\,, \tag{25}$$

$$|\tilde{\rho}\rangle\!\rangle = \sum_j \langle\!\langle E|\,F_j\,|\rho\rangle\!\rangle\,|j\rangle\!\rangle \tag{26}$$

$$\left\langle\!\langle\tilde{E}\right| = \sum_k \langle\!\langle E|\,F_k\,|\rho\rangle\!\rangle\,\langle\!\langle k|, \tag{27}$$

and then construct $\{|\hat{\rho}\rangle\!\rangle, \langle\!\langle\hat{E}|, \{\hat{G}_i\}\}$ as above.

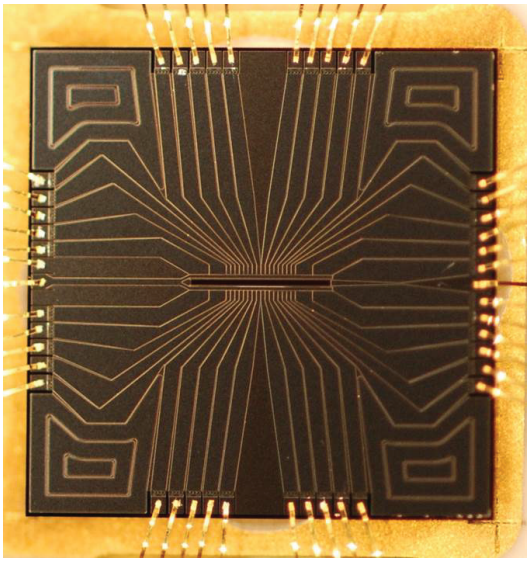

FIG. 2: **Surface Electrode Trap.** Linear surface electrode ion trap used for the experiment. The trap has a central through substrate slot. Neutral ytterbium vapor reaches the trapping volume through the slot from the back of the chip. For these experiments a single $^{171}Yb^{+}$ is trapped in the center of the trap.

$\tilde{\mathbb{1}}$ may not be invertible – but if and only if either A or B is rank-deficient. This occurs only if either the set $\{\rho_k\}$ or the set $\{E_j\}$ fails to span $B(\mathcal{H})$ – i.e., they are informationally incomplete. This is easily diagnosed by simply checking the rank of $\tilde{\mathbb{1}}$. If it occurs, we replace some of the $\{F_k\}$ with alternative sequences that *do* produce informationally complete sets. If this consistently fails to fix the rank-deficiency, it indicates that the gate set is not sufficiently universal to generate an informationally complete set, which requires hardware-level intervention.

If $\tilde{\mathbb{1}}$ is full-rank, but has small eigenvalues, the experiments are marginally informationally complete. Small statistical fluctuations in the observed frequencies will be amplified by the inversion. This too can be fixed by adding more sequences $\{F_k\}$ to the mix and casting out the least useful ones [i.e., the ones whose removal maximizes $\lambda_{\min}(\tilde{\mathbb{1}})$].

IV. IMPLEMENTATION OF GST ON A TRAPPED-ION QUBIT

Trapped ions are among the most reliable qubits available today; up to 14 qubits have been addressed in a single trap [23], while logic gates on single qubits have been performed with sustained failure probabilities of around 2×10^{-5} [5]. In order to scale these demonstrations to the large number of qubits needed for quantum information processing protocols it is crucial to use micro-fabricated trap structures. Micro-fabrication enables the fabrication of extended segmented traps that provide the abil-

ity to use multiple trapping sites and to shuttle ions between different locations. Sandia National Laboratories uses state of the art silicon fabrication technology to produce sophisticated and highly optimized surface electrode traps for use in quantum information processing experiments. We used a Sandia surface trap to demonstrate GST and our coherent qubit manipulation capabilities.

We trap a single $^{171}\mathrm{Yb}^{+}$ ion in a linear surface ion trap, shown in Fig. 2 [24, 25], by photoionizing neutral ytterbium vapor that reaches the trapping volume through a slot from the back of the surface trap chip. The qubit is encoded in the $|F=0, m_F=0\rangle$ and $|F=1, m_F=0\rangle$ hyperfine clock states of the $^2S_{1/2}$ ground state of $^{171}\mathrm{Yb}^{+}$ which are labeled $|1\rangle$ and $|0\rangle$, respectively.

Standard laser cooling techniques are applied to Doppler cool the ion and prepare it in the $|0\rangle$ state. The quantum state is read out via standard fluorescence state detection[26]. Microwave radiation resonant with the 12.6428 MHz separation of the qubit levels is used to control the qubit. For a π-pulse microwave radiation with a square envelope is applied for approximately 58μs.

We used GST to characterize a set of four gates that generate the full set of single-qubit Clifford gates in this system. Because the primary purpose of this experiment was to evaluate and demonstrate GST, we did not attempt to minimize errors in our gate set. However, our analysis showed that the gates are extremely accurate – enough that even standard tomography would have worked fairly well (although it is only thanks to the robust GST framework that we can say this with confidence!)

We implemented an alphabet of four quantum operations ($\{G_1 \dots G_4\}$), aiming at the target set

$$T_1 = \mathbb{1} \quad (28)$$
$$T_2 = e^{i(\pi/4)\sigma_x} \quad (29)$$
$$T_3 = e^{i(\pi/4)\sigma_y} \quad (30)$$
$$T_4 = e^{i(\pi/2)\sigma_x}. \quad (31)$$

Our target initial state was $\rho_{\mathrm{ideal}} = |1\rangle\langle 1|$, and our target measurement was $\mathcal{M}_{\mathrm{ideal}} = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$.

We performed two kinds of gate sequences to gather data: *training* and *testing*. The training sequences generated data that was used to generate tomographic estimates. The testing sequence data (discussed below) were kept hidden until after the estimates had been generated, and were then used to objectively "score" the four different kinds of tomography, by evaluating how well they predicted the testing data. In order to minimize the effect of systematic drift during the several hours required to take all this data, we interleaved training and testing experiments. We also recalibrated the gate pulse amplitude periodically.

Our training sequences were designed around the demands of LGST. We chose the simplest possible fiducial sequences, $F_k = G_k$ for $k = 1 \dots 4$. For each of the five operations $X \in \{\mathbb{1}, G_1 \dots G_4\}$, we performed 16 distinct experiments of the form

$$\langle\langle E| F_i X F_j |\rho\rangle\rangle$$

to estimate the 4×4 matrix $\tilde{X}$. (For $X = \mathbb{1}$, the experiments were of the form $\langle\langle E| F_i F_j |\rho\rangle\rangle$). An additional 4 experiments involving just one gate – $\langle\langle E| F_i |\rho\rangle\rangle$ – were performed to enable inference of ρ and E. Each of these 84 experiments was repeated 1900 times to obtain statistics (each repetition yielded a single binary result, depending on whether the ion fluoresced). The formulae from Sec. III were then used to calculate LGST estimates of ρ, E, and $\hat{G}_1 \dots \hat{G}_4$. Since these estimates are only defined up to a gauge (see Sec. II B), we then used a numerical search to find the gauge transformation

$$\hat{G}_k \to M G_k M^{-1}$$

that minimized the RMS discrepancy,

$$\sum_k \mathrm{Tr}[(G_k - T_k)^2],$$

between the estimated $\hat{G}_k$ and the target gates T_k. The resulting *linear* GST estimates, represented as 4×4 superoperators in the Pauli basis, and presented adjacent to the target gates, are given in Table I.

	LGST estimate ($\hat{G}_k$)	Target (T_k)
ρ	$\begin{pmatrix} 0.0099 & 0.0104+0.0007i \\ h.c. & 0.9901 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$
E	$\begin{pmatrix} 0.9879 & 0.0182-0.0023i \\ h.c. & 0.0121 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$
G_1	$\begin{pmatrix} 0.9977 & -0.0219 & -0.0204 & 0.0024 \\ -0.0152 & 0.9657 & 0.017 & 0.0291 \\ 0.0031 & 0.0627 & 1.0172 & 0.0335 \\ 0.001 & 0.0065 & 0.0335 & 0.9915 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
G_2	$\begin{pmatrix} 0.9974 & -0.048 & -0.0304 & 0.0161 \\ -0.0077 & 0.9538 & -0.0033 & -0.0045 \\ -0.0113 & 0.0332 & 0.0066 & -1.0044 \\ -0.0029 & 0.0042 & 1.0099 & 0.0284 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
G_3	$\begin{pmatrix} 0.9923 & -0.0163 & -0.0066 & 0.001 \\ -0.0049 & -0.0087 & -0.0087 & 0.9839 \\ 0.0124 & -0.0082 & 1.0136 & -0.0017 \\ -0.0074 & -0.9797 & 0.0043 & 0.0025 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$
G_4	$\begin{pmatrix} 0.9991 & -0.0291 & 0.0028 & 0.0194 \\ 0.0096 & 0.9796 & -0.0049 & 0.0013 \\ 0.0083 & -0.0211 & -1.0494 & -0.0632 \\ -0.0091 & -0.0123 & -0.0427 & -1.0012 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$

TABLE I: **Results of LGST**. This table shows the LGST estimate of our trapped-ion qubit gates, based on 84 distinct experiments (gate sequences), each repeated 1900 times. The intended target gates are shown on the right. Estimates (obtained on the left) were obtained using the LGST analysis procedure given in Section III, then gauge-optimized numerically, by applying similarity transformations to all gates, to match the target gates as closely as possible.

	ML estimate (short dataset)	ML estimate (long dataset)	Target gates
ρ	$\begin{pmatrix} 0.0099 & 0.0077-0.0046i \\ h.c. & 0.9901 \end{pmatrix}$	$\begin{pmatrix} 0.0092 & -0.0017+0.0088i \\ h.c. & 0.9908 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$
E	$\begin{pmatrix} 0.9911 & 0.0166-0.0006i \\ h.c. & 0.0089 \end{pmatrix}$	$\begin{pmatrix} 0.988 & 0.0019+0.0089i \\ h.c. & 0.012 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$
G_1	$\begin{pmatrix} 1.0019 & -0.0128 & -0.0198 & -0.0002 \\ -0.0066 & 0.9775 & -0.0118 & 0.0122 \\ 0.0041 & 0.0842 & 1.0138 & 0.0073 \\ -0.0035 & -0.013 & 0.0075 & 0.9969 \end{pmatrix}$	$\begin{pmatrix} 1.0001 & -0 & 0.0003 & 0.0001 \\ 0.0001 & 0.9994 & -0.0003 & -0 \\ -0.0001 & 0.0006 & 0.999 & -0.0003 \\ -0 & -0.0001 & 0.0002 & 0.9998 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
G_2	$\begin{pmatrix} 1.0017 & -0.0276 & -0.0276 & -0.0048 \\ -0.0193 & 0.9582 & -0.0076 & -0.0127 \\ -0.0134 & 0.043 & 0.0082 & -0.9987 \\ -0.0072 & 0.002 & 1.0069 & 0.0192 \end{pmatrix}$	$\begin{pmatrix} 1 & -0.0001 & -0.0045 & -0.0005 \\ 0 & 0.9994 & -0.006 & -0.0018 \\ -0.005 & -0.0112 & -0.0064 & -0.9991 \\ 0.0006 & 0.0063 & 0.9993 & 0.0143 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
G_3	$\begin{pmatrix} 0.99 & -0.0114 & 0.0083 & 0.0044 \\ -0.0082 & -0.0141 & -0.0045 & 0.9892 \\ 0.0121 & -0.0044 & 1.0056 & -0.0059 \\ -0.0001 & -0.9848 & 0.0017 & -0.0016 \end{pmatrix}$	$\begin{pmatrix} 1.0001 & 0.0033 & 0.0001 & 0.0049 \\ 0.0033 & -0.0001 & -0.0005 & 0.9992 \\ -0.0002 & -0.0024 & 0.9995 & -0.0161 \\ -0.0019 & -0.9989 & 0.0179 & 0.0085 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$
G_4	$\begin{pmatrix} 0.9983 & -0.0217 & 0.0127 & 0.0142 \\ -0.0039 & 0.9745 & 0.0034 & 0.0077 \\ -0.0004 & -0.0145 & -1.0473 & -0.0323 \\ -0.014 & -0.0167 & -0.0072 & -1.0024 \end{pmatrix}$	$\begin{pmatrix} 1.0001 & -0 & 0.0062 & 0.0028 \\ -0 & 0.9997 & 0.0127 & 0.0022 \\ 0.0066 & 0.0164 & -0.9976 & 0.0065 \\ -0.004 & -0.0004 & -0.0066 & -0.9981 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$

TABLE II: **Maximum likelihood refinements of LGST gates**. This table shows maximum likelihood (ML) estimates of the gate set. Column 2 shows the results of ML estimation on the "short" dataset of 85 sequences (the LGST sequences and the SPAM sequence $\langle\langle E|\rho\rangle\rangle$). Column 3 shows the results of ML estimation on the "long" dataset of 1066 sequences described in the text. Column 4 shows the target gates that we intended to implement.

V. IMPROVING GST WITH MAXIMUM LIKELIHOOD

Linear inversion tomography has been largely superseded by maximum likelihood estimation (MLE), for multiple reasons. Linear inversion and MLE coincide when the data are informationally complete (rather than overcomplete) and the linear inversion estimate doesn't violate positivity constraints. But the ML method can easily take account of constraints *and* can reconcile overcomplete data efficiently, both of which are essentially impossible for linear inversion (least squares optimization is properly seen as an approximation to MLE, rather than a generalization of linear inversion).

For gate-set tomography, a third quality is even more important: maximum likelihood is easily adapted to *nonlinear* data – i.e., the results of experiments in which the directly inferrable probabilities are not linear functions of the parameters. Such experiments are natural in gate set tomography, and promise great improvements in accuracy. Probabilities involving G_k^n are roughly n times more sensitive to variations in G_k than probabilities depending linearly on G_k. However, nonlinear data poses a danger; the likelihood function $\mathcal{L}(G)$ need not be unimodal or have convex level sets, which means that maximizing a generic GST likelihood function is not a convex problem, and may be NP-hard.

Fortunately, LGST provides a simple solution to this problem. The LGST estimate is typically not optimal, but it is necessarily *close* to the point of maximum likelihood, and we can reasonably expect that a gradient ascent algorithm starting from the LGST estimate will find the global maximum of $\mathcal{L}(G)$. So ML acts as a turbocharger for GST, relying critically on the LGST estimate to provide a good starting estimate, and then refining it to incorporate the nonlinear and overcomplete data that are critical for achieving high accuracy.

We use the standard Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [27], as implemented in *Scipy* [28], to minimize the likelihood function. With no particular attention paid to optimization, run times ranged from a few seconds to several hours (depending on the complexity of the data) on a typical laptop computer. The LGST estimate often does not predict positive probabilities for the training data, which means that the log-likelihood is technically undefined when it is chosen as a starting point. To address this, we first find an in-bounds starting point by using the Nelder-Mead downhill simplex method [29] to find the nearest valid gate set (in terms of Euclidian distance). We then use this point as the initial value for our BFGS optimization routine.

We found ML estimates for two datasets: (1) the $84 = 4+16\times5$ LGST sequences of length ≤ 3 only, and (2) a set of 1066 distinct sequences (each repeated 1900 times) corresponding to LGST on: $\{G_k\}$, $\{G_k^2\}$, $\{G_k^4\}$,... $\{G_k^{128}\}$. These choices of experiments were somewhat arbitrary;

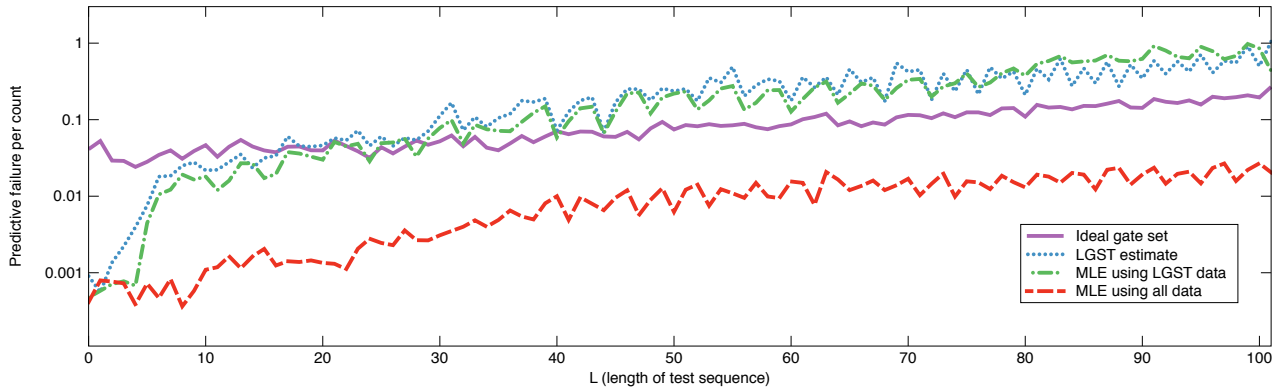

FIG. 3: **Average score versus test sequence length.** This figure shows the logarithmic *predictive score* achieved by four different estimates on 1000 distinct testing experiments. Each experiment corresponded to a partial sequence comprising the first L gates from one of 10 sequences of 100 gates (see Figs. 4-5), and was repeated 950 times. The vertical axis shows the average (per count) logarithmic score achieved by each of five estimates as a function of L (averaged over the 10 partial sequences of length L). Notably, the bare LGST estimate outperforms target gates when used to predict short sequences, but becomes rapidly very inaccurate for sequences longer than $L = 5$. The ML estimate incorporating data from long training strings is far more accurate for $L > 5$, achieving a per-count score of ≈ 0.01 even on strings of length $L = 100$. For reference, an estimate that predicted $p = \frac{1}{2}$ for every experiment would suffer a score of roughly $\frac{\ln(2)}{2} \approx 0.35$ per count.

we have no reason to believe that these sequences will provide better (or worse) accuracy than any other sequences of various lengths. We refer to these as the "short" and "long" datasets. The MLE estimates are given in Table II.

VI. QUANTIFYING ACCURACY OBJECTIVELY WITH SCORED TESTS

Tomography is not the *end* of a science experiment; it is the middle. The tomographic estimate is a theory; it needs to be tested to determine how well it predicts further experiments. We cannot evaluate the theory based on how well it fits past ("training") data, since its parameters were chosen specifically to fit them. Tomographic estimates are traditionally scored using some concept of fidelity, but this is always problematic. First, the whole point of tomography is to characterize unknown quantities, so we don't have a "true" state/process with which to evaluate fidelity. Second, the gauge degree of freedom in gate set tomography makes it unclear how to calculate or interpret standard quantities like entanglement fidelity or diamond norm.

We therefore introduce a novel and very simple method for evaluating tomographic estimates. We perform a set of "testing" experiments – sequences of gates that were not performed in the tomographic phase – and score our tomographic estimates based on how well they predict the results. The scoring is based entirely on observable probabilities, which are explicitly gauge-invariant. Of course, there are many ways to compare (predicted) probabilities to (empirical) frequencies. The *log scoring rule* is particularly simple and well-motivated.

For each testing experiment S_j, we use the tomographic estimate[s] to assign probabilities to the outcome (before it is revealed). Each of our experiments has 2 outcomes, which we label + and −, so the predicted probabilities are $\{p_+, p_-\}$. When the outcome (call it "b") is revealed, we increment the estimate's score by the negative logarithm of $Pr(b)$:

$$\text{score} \to \text{score} - \log(p_b).$$

When all the testing data are evaluated, this leads to a total score of

$$\text{score}_0 = -\sum_j n_+(j)\log[p_+(j)] + n_-(j)\log[p_-(j)],$$

where lower scores are better. We then renormalize the score by subtracting off the minimum score that *any* prediction could conceivably achieve (because some of the score is due to the entropy of the data itself):

$$\begin{aligned}\text{score} &= \sum_j n_+(j)\log\left[\frac{n_+(j)}{n_+(j)+n_-(j)}\right] \\ &+ \sum_j n_-(j)\log\left[\frac{n_-(j)}{n_+(j)+n_-(j)}\right] \\ &- \sum_j n_+(j)\log[p_+(j)] + n_-(j)\log[p_-(j)].\end{aligned}$$

This score is in fact (1) the relative entropy between the predicted probabilities and the empirical frequencies, and (2) the loglikelihood of the tomographic estimate given the testing data.

However, while the logarithmic score is very well-motivated, it penalizes nonpositive probability estimates rather dramatically – if $p = 0$, then the penalty is

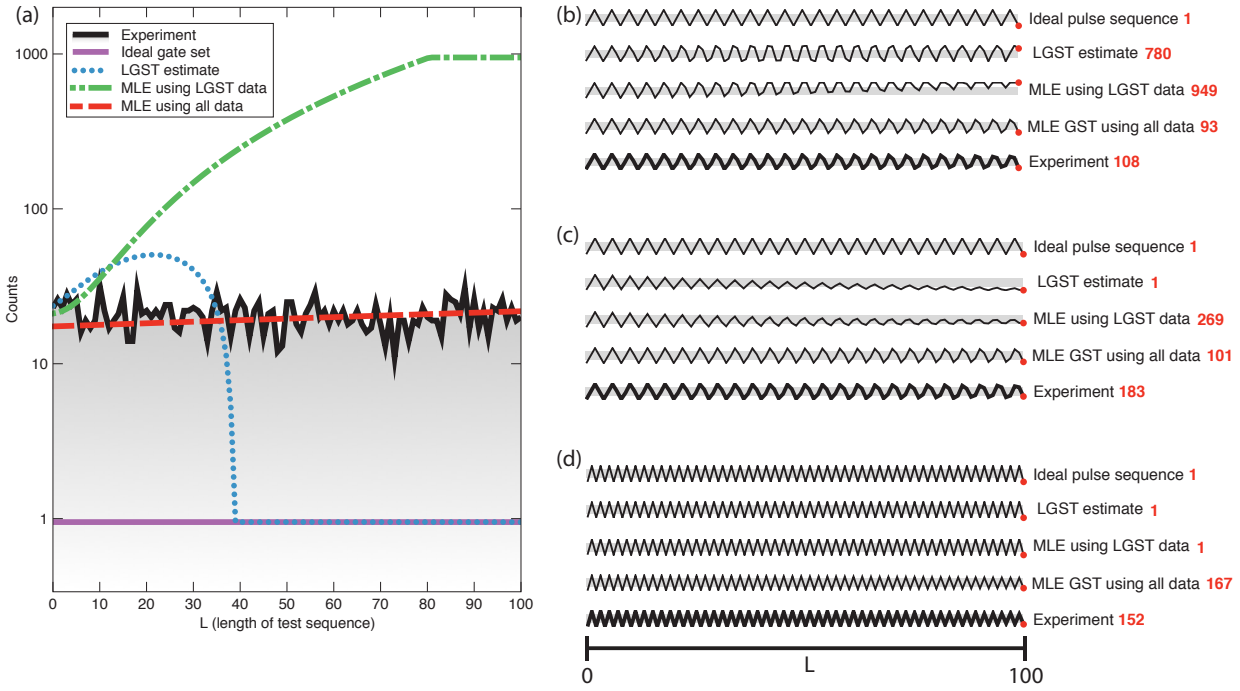

FIG. 4: **Rabi oscillations – prediction vs. data**. Each panel shows (i) observed counts and (ii) predicted counts, for one of ten series of testing experiments. Each testing series was based on a particular sequence of 100 consecutive gates, which was used to define a series of 101 experiments corresponding to *partial* sequences of length L (horizontal axes). Each partial sequence was repeated 950 times to obtain the displayed count statistics (vertical axes). The sequences shown in this figure are: **(a)** 100 consecutive $\mathbb{1}$ gates; **(b)** 100 consecutive $X_{\pi/2}$ gates; **(c)** 100 consecutive $Y_{\pi/2}$ gates; and **(d)** 100 consecutive X_{π} gates. Each plot thus represents a Rabi oscillation experiment (albeit with discrete gates rather than continuous Hamiltonian evolution), and compares the observed counts to the predictions of ML GST estimates obtained from both short and long training datasets. In panels **(b)-(d)**, the number after each label indicates the counts at the point at the end of each trace. The grey shaded bars represent the middle 50%. of possible counts.

$-\log(0) = \infty$, and if $p < 0$, the whole formalism fails. Although negative probabilities should never be predicted, we did not impose positivity in this work, so some of our estimates (especially naive tomography and LGST) do predict negative probabilities. We deal with this in a simple and tolerably well-motivated way: whenever an estimate predicts $p < \epsilon$ for some small threshold ϵ, we truncate that probability to ϵ. In the data reported here, we chose $\epsilon = 10^{-3}$ [approximately $1/N$, where each training sequence was measured $N = 1900$ times; after N observations, the lowest probability that can reasonably be reported is $O(1/N)$], and verified that varying ϵ does not qualitatively change the results.

We scored and compared four different estimates, each of which can be used to predict the testing data:

1. The target gates themselves,
2. LGST using only the $4 + 5 \times 16 = 84$ training sequences discussed above,
3. Maximum likelihood GST on the "short" 85-sequence dataset comprising 84 LGST sequences and the "SPAM" experiment $\langle\langle E|\rho\rangle\rangle$,
4. Maximum likelihood GST using a much richer "long" dataset, described below.

Our "long" dataset was intended to probe the use of long gate sequences: (1) their utility for improving accuracy; and (2) the feasibility of GST estimation for such data. We included the 85 sequences in the "short" dataset, and added $448 = 7 \times 4 \times 16$ additional sequences of the form $\langle\langle E| F_i G_k^p F_j |\rho\rangle\rangle$ for $p = 2, 4, 8, 16, 32, 64, 128$. That is, we did the experiments necessary for LGST on $\{G_k^p\}$, although we only analyzed this data using MLE. Finally, for each of these 533 experiments, we *also* performed a corresponding experiment in which we added a single $G_4 \approx e^{i(\pi/2)\sigma_x}$ gate at the end, so that we could probe the bright-vs-dark asymmetry of the measurement[1]. The "long" dataset thus contains a total of 1066 sequences, each repeated 1900 times, for a total of $\approx 2 \times 10^6$ measurements.

[1] This turned out to not actually be necessary – GST is completely self-calibrating, and can extract this asymmetry from any complete set of training sequences.

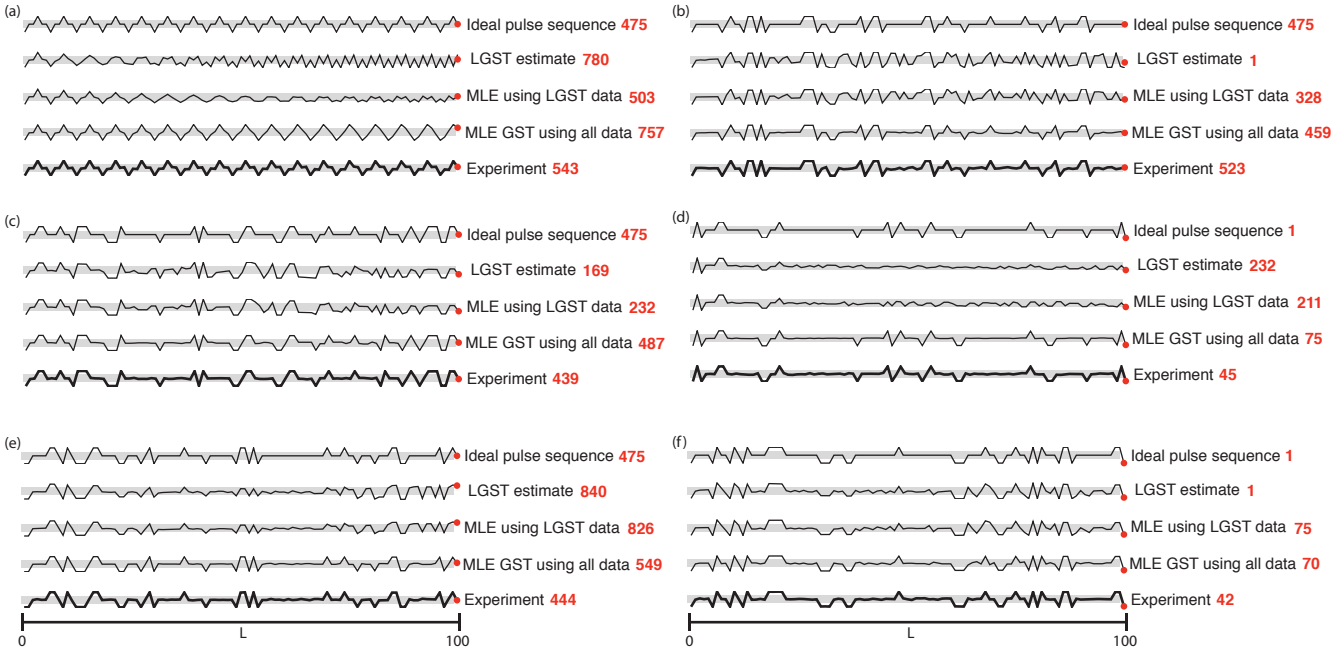

FIG. 5: **Generalized Rabi oscillations – prediction vs. data**. These plots show essentially the same sort of results as Figure 4, but on non-uniform strings of gates. Panel **(a)** shows a sequence of 100 alternating $X_{\pi/2}$ and $Y_{\pi/2}$ gates, while Panels **(b-f)** show the results of five randomly chosen gate sequences (similar to those that would be involved in randomized benchmarking). Each plot thus represents a "generalized Rabi oscillation experiment", in that we are evaluating the accuracy with which a given estimate predicts an evolution through quantum state space, but that evolution is not the orbit of a Hamiltonian. In each panel, he number after each label indicates the counts at the point at the end of each trace. The grey shaded bars represent the middle 50%. of possible counts.

The main results are displayed in Figure 3, which shows the estimates' score-per-count (note: lower scores are better), averaged over all 10 test sequences of length L, as a function of L. Shorter sequences are easier to predict, and all estimates' scores increase with L.

All three tomographic estimates predict very short ($L \leq 5$) fairly well. The target gates themselves fail to predict even short sequences well, although most of this predictive failure seems to be due to SPAM error rather than errors in the gates. Maximum likelihood methods are more accurate than LGST even for strings of length $L = 4, 5$, but the most dramatic difference is between the "Long ML" estimates, which were trained on long sequences, and all the others. For reference, we note that an estimator that simply predicted $p = \frac{1}{2}$ for every count would achieve a score-per-count of approximately $\frac{\ln 2}{2} \approx 0.35$. The best ML estimate still achieves a score-per-count of ~ 0.02 at $L = 100$, indicating a very high degree of predictive power even for long test sequences.

The LGST estimate works well on strings of the same length as its training data, but degrades rapidly beyond $L = 3$. This is not a major concern, however. LGST's critical role is to get *close enough* to provide a good seed for ML methods (which it does admirably), not to provide an optimal estimate. We suspect that LGST's accuracy could be improved quite a bit by using overcomplete data and appropriately weighted least-squares fitting.

VII. CONCLUSIONS

Continued development of QIP technology – memory qubits, logic gates, state preparations, and measurements – depends critically on reliable characterization protocols, which cannot rely on precalibrated reference frames that are not available in most technologies. Gate set tomography is, to our knowledge, the first completely reliable framework and protocol for characterizing quantum logic gates. By using LGST as a first stage to obtain a closed-form approximation to the true gates, the GST protocol can ensure robustness against local maxima in the likelihood function – yet take full advantage of maximum likelihood (or any other well-motivated statistical method) to achieve high accuracy.

Our experimental demonstration illustrates GST's ability, *and* demonstrates that it is practically feasible. Moreover, out of necessity, we have introduced and demonstrated a novel and (we think) very useful method for objectively testing how good a tomographic estimate is. Unlike all the previous work of which we are aware, this scoring protocol doesn't measure how well the tomographic estimate agrees with the target goal (which might be incorrect) or with another tomographic estimate (which might be biased in the same way as this one). Instead, it evaluates how well the estimate does its fundamental job – predicting future data. Our results

not only illustrate the scoring protocol, but also show that our GST estimates are quite good predictors.

We do not expect that gate set tomography will be another kind of tomography, standing shoulder to shoulder with state tomography, process tomography, and measurement tomography. It is intended to *replace* them – to be, as one of us has said in public, "One tomography to rule them all." This is out of necessity: state tomography requires well-calibrated measurements, measurement tomography requires well-calibrated states, and gate tomography requires both – yet in practice, states and measurements are only as well calibrated as the gates that prepare them! This is a vicious circle. GST cuts that Gordian knot by (1) estimating everything self-consistently, and (2) identifying the *gates* as the critical element. Gates are central because they can be applied multiple times in a single experiment (unlike state preparations and measurements, which can appear only once per experiment), and this allows us to generate combinatorially many (2^L) distinct observable probabilities using only 2 distinct gates (and thus without adding any extra parameters to be estimated).

The necessary price paid for this is the appearance of the $SL(d^2)$ gauge. This gauge is arguably the single most intriguing and pernicious aspect of GST. It is clearly fundamental to black-box descriptions of quantum devices, and therefore seems to be fundamental to QIP. Yet it interacts very badly with complete positivity, and we do not yet know how to represent gate sets in an efficient and gauge-invariant way – nor how to compute a gauge-invariant measure of fidelity between two gate sets (e.g., a target and an estimate). Of all the many aspects of GST that cry out for further research and development, the gauge – its relationship with conventional descriptions of circuit QIP, and how it can be tamed – seems the most worthy of urgent study.

Acknowledgments

RBK is grateful for helpful conversations with Steve Flammia, Stephen Bartlett, Jay Gambetta, Seth Merkel, and Cyril Stark. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

[1] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Phys. Rev. A **87**, 062119 (2013).
[2] M. Takahashi, S. D. Bartlett, and A. C. Doherty, Phys. Rev. A **88**, 022120 (2013).
[3] J. Medford, J. Beil, J. M. Taylor, S. D. Bartlett, A. C. Doherty, E. I. Rashba, D. P. Divincenzo, H. Lu, A. C. Gossard, and C. M. Marcus, Nature Nanotechnology **8**, 654 (2013).
[4] D. Mahler, L. A. Rozema, A. Darabi, C. Ferrie, R. Blume-Kohout, and A. Steinberg, arXiv preprint arXiv:1303.0436 (2013).
[5] K. R. Brown, A. C. Wilson, Y. Colombe, C. Ospelkaus, A. M. Meier, E. Knill, D. Leibfried, and D. J. Wineland, Phys. Rev. A **84**, 030303 (2011).
[6] D. Mogilevtsev, J. Řeháček, and Z. Hradil, New Journal of Physics **14**, 095001 (2012).
[7] A. Brańczyk, D. H. Mahler, L. A. Rozema, A. Darabi, A. M. Steinberg, and D. F. James, New Journal of Physics **14**, 085003 (2012).
[8] C. Stark, arXiv preprint arXiv:1209.5737 (2012).
[9] C. Stark, arXiv preprint arXiv:1209.6499 (2012).
[10] C. Stark, arXiv preprint arXiv:1210.1105 (2012).
[11] K. Vogel and H. Risken, Phys. Rev. A **40**, 2847 (1989).
[12] D. Smithey, M. Beck, M. Raymer, and A. Faridani, Physical review letters **70**, 1244 (1993).
[13] M. Paris and J. Řeháček, *Quantum state estimation*, vol. 649 (Springer, 2004).
[14] I. L. Chuang and M. Nielsen, Journal of Modern Optics **44**, 2455 (1997).
[15] Z. Hradil, Physical Review A **55**, 1561 (1997).
[16] R. Blume-Kohout, New J. Phys. **12**, 043034 (2010).
[17] R. Blume-Kohout, Phys. Rev. Lett. **105**, 200504 (2010).
[18] J. A. Smolin, J. M. Gambetta, and G. Smith, Phys. Rev. Lett. **108**, 070502 (2012).
[19] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Nature Communications **1**, 149 (2010).
[20] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Phys. Rev. Lett. **105**, 150401 (2010), URL `http://link.aps.org/doi/10.1103/PhysRevLett.105.150401`.
[21] W. van Dam, F. Magniez, M. Mosca, and M. Santha, in *Proceedings of the thirty-second annual ACM symposium on Theory of computing* (ACM, 2000), pp. 688–696.
[22] R. Blume-Kohout and et. al., in preparation (2013).
[23] T. Monz, P. Schindler, J. T. Barreiro, M. Chwalla, D. Nigg, W. A. Coish, M. Harlander, W. Hänsel, M. Hennrich, and R. Blatt, Phys. Rev. Lett. **106**, 130506 (2011).
[24] D. Stick, K. M. Fortier, R. Haltli, C. Highstrete, D. L. Moehring, C. Tigges, and M. G. Blain, arXiv preprint arXiv:1008.0990 (2010).
[25] E. Mount, S.-Y. Baek, M. Blain, D. Stick, D. Gaultney, S. Crain, R. Noek, T. Kim, P. Maunz, and J. Kim, arXiv preprint arXiv:1306.1269 (2013).
[26] S. Olmschenk, K. C. Younge, D. L. Moehring, D. N. Matsukevich, P. Maunz, and C. Monroe, Physical Review A **76**, 052314 (2007).
[27] J. Nocedal and S. Wright, *Numerical optimization*, Operations research and financial engineering (Springer, New York, 2006).
[28] E. Jones, T. Oliphant, P. Peterson, et al., *SciPy: Open source scientific tools for Python* (2001–), URL `http://www.scipy.org/`.
[29] J. Nelder and R. Mead, Computer Journal **7**, 308 (1965).