Operational, gauge-free quantum tomography

Olivia Di Matteo[1,2,3], John Gamble[4], Chris Granade[4], Kenneth Rudinger[5], and Nathan Wiebe[4,6,7]

[1] TRIUMF, Vancouver, British Columbia, Canada V6T2A3

[2] Department of Physics and Astronomy, University of Waterloo, Waterloo, ON, Canada

[3] Institute for Quantum Computing, University of Waterloo, Waterloo, ON, Canada

[4] Microsoft Research, Quantum Architectures and Computation Group, Redmond, Washington 98052, USA

[5] Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

[6] Department of Physics, University of Washington, Seattle, WA 98195, USA

[7] Pacific Northwest National Laboratory, Richland, WA 99352, USA

authors in alphabetical order by last name

As increasingly impressive quantum information processors are realized in laboratories around the world, robust and reliable characterization of these devices is now more urgent than ever. These diagnostics can take many forms, but one of the most popular categories is *tomography*, where an underlying parameterized model is proposed for a device and inferred by experiments. Here, we introduce and implement efficient operational tomography, which uses experimental observables as these model parameters. This addresses a problem of ambiguity in representation that arises in current tomographic approaches (the *gauge problem*). Solving the gauge problem enables us to efficiently implement operational tomography in a Bayesian framework computationally, and hence gives us a natural way to include prior information and discuss uncertainty in fit parameters. We demonstrate this new tomography in a variety of different experimentally-relevant scenarios, including standard process tomography, Ramsey interferometry, randomized benchmarking, and gate set tomography.

Contents

1 Introduction

Quantum computing offers the potential for significant advantages across a wide range of important problems. Establishing a rigorous understanding of the costs involved in producing enterprise-scale quantum computers is a critical part of current decision making. This need has driven efforts to more precisely estimate the costs of different algorithms across different applications, such as in quantum chemistry simulations [1]. However, these resource estimations depend critically on the *quality* of the qubits used, *i.e.*, the accuracy with which one can perform quantum gates and measurements. The collection of procedures used for detecting and debugging faulty operations on quantum computers is known as quantum characterization, verification, and validation (QCVV). Through QCVV, scientists and engineers working on quantum hardware can hope to diagnose errors and certify performance, in turn improving qubit design and operation.

One goal of QCVV is to learn what actually happens when we attempt to apply a target unitary operator U, a procedure broadly known as *quantum tomography*. Using the language of open quantum systems, we can hypothesize that there is some *channel* Λ that, if we knew it, would allow us to predict what happens when we apply U to any state. The problem then becomes determining how should we best learn Λ given experimental evidence from our quantum device. The tomography problem has been approached in a wide variety of ways [2–12]. The various procedures generally learn Λ by (repeatedly) preparing a variety of input states $\{\rho_i\}$, sending each through the application of U, and then measuring a variety of effects $\{E_j\}$. In some cases, the use of auxiliary qubits as a reference can eliminate the need to vary over input states [13]. However, these latter approaches are mainly useful for reasoning mathematically about tomography [4] and offer limited experimental applicability. Hence, we will focus on the more typical case in this work.

While valid, this rests critically on the assumption that we know what each of $\{\rho_i\}$ and $\{E_j\}$ are. In practice, each state ρ_i and each measurement E_j may be subject to its own physical errors, and these may in turn be objects that we would like to learn. Worse still, we often prepare states by performing a particular privileged state preparation procedure $\rho_\star$, and then applying unitary evolution operators $\{V_i\}$ to obtain $\rho_i := V_i \rho_\star V_i^\dagger$. Similarly, measurements are often effected by transforming a particular privileged measurement under unitary evolution.

Once we include the experimental consideration that the channels we would like to study are the same ones that we use to prepare and measure our devices, we are forced to ensure that we learn the channels describing our system in a self-consistent manner. We cannot learn a channel Λ entirely independently of the experimental context in which Λ occurs, but must describe that channel such that we can predict its action in a larger experiment. This self-consistency requirement then forces us to face another difficulty: we can always transform the entire description of an experiment in a consistent fashion, such that there is *no* observable difference whatsoever. For instance, the states $|0\rangle$ and $|1\rangle$ are in essence just labels for two levels of a quantum system; there is no observable impact to our calling them $|\heartsuit\rangle$ and $|\diamondsuit\rangle$, $|\sharp\rangle$ and $|\flat\rangle$, or even $|1\rangle$ and $|0\rangle$.

That we can rename $|0\rangle$ to $|1\rangle$ and vice versa illustrates one way to formally describe the challenge imposed by self-consistency. In particular, if we perform an experiment whose outcomes are described by Born's rule as $\mathrm{Tr}(E\Lambda[\rho])$, then for any unitary map U the experiment $\mathrm{Tr}(UEU^\dagger \cdot U\Lambda[U^\dagger \cdot U\rho U^\dagger \cdot U]U^\dagger)$ has the exact same outcome distribution, and thus cannot be distinguished from our original description. That is, we cannot decide if we have (ρ, E, Λ) or if we have $(U\rho U^\dagger, UEU^\dagger, U\Lambda U^\dagger)$.

Taking a step back, something seemingly ridiculous has happened: we asked merely for a description of how one component of our quantum device operates, and arrived at a seemingly fundamental limit to what knowledge we can ever gather about our device. After all, $U\rho U^\dagger$ and ρ seem to be very different preparation procedures! Recently, gate set tomography (GST) [14, 15] has been used as a means to solve this conundrum by explicitly including the effects of this apparent ambiguity into estimation procedures. With GST, we perform inference on the entire gate set, state preparation, and measurement procedure based on empirical frequencies from repeated experiments. This inference procedure can be quite sophisticated in practice, with carefully designed experiments to tease out very slight channel imperfections. Over the past several years, GST has been demonstrated experimentally on a wide variety of platforms [15–28], predominately using the software package pyGSTi [29, 30].

Of course, gate set tomography also has drawbacks, suffering from a conceptual difficulty known as the *gauge problem* [17, 31, 32]. While gauge-invariant scoring metrics have been used in the past [15] (as we do in the present work), we note that the underlying representation of the gate set used to carry out GST is gauge-variant. Specifically, GST eschews any notion of a fixed reference frame in favor of a *gauge group* that specifies how to transform a valid estimate of an error model into a family of related error models which give identical experimental predictions. While such gauge transformations do not impact the predictions made by such a model, they *do* impact the particular channels reported at the end of the inference procedure, and some commonly reported metrics on channels are not gauge-invariant. In practice, one gauge-fixes resulting channels to some external reference frame, but this procedure requires nonlinear optimization whose global convergence is not guaranteed. Finally, a procedure for systematically including prior information in GST has not yet been put forward, which could potentially result in massive savings if developed.

In this paper, we introduce *operational quantum tomography* (OQT), which is a general (operational) framework that allows us to reason about a host of different tomographic procedures (including GST) in a manifestly gauge-independent manner. In addition to resolving the gauge problem, OQT allows us to naturally include prior information in GST within Bayesian inference, which was computationally prohibitive previously due to the the gauge fixing procedure.

OQT is enabled by using a new, manifestly gauge-invariant, representation of our gate set. This representation is inspired by linear-inversion gate set tomography [15]; we term this the *operational representation*. After introducing the operational representation and explaining how it resolves the gauge problem, we discuss how to implement OQT numerically within a Bayesian framework while including prior information. We then detail the performance of this technique by tracking prediction loss, a useful and gauge-invariant measure of the quality of our ability to predict the outcome of future experiments, across a suite of experimentally relevant problems: Ramsey interferometry, quantum process tomography, and randomized benchmarking. We close by showing how dynamics of quantum systems may, in general, be described using the operational representation.

2 Gate set tomography and the operational representation

2.1 GST formalism

As described in the introduction, quantum state and process tomography make strong assumptions about our ability to perform state preparation and measurement (SPAM). Tomographic reconstructions of states and processes that are made assuming perfect SPAM will be inconsistent with the true, noisy operations. A key advantage of GST is that it produces *self-consistent* estimates by simultaneously characterizing SPAM along with other processes.

Here we briefly review GST, following Refs. [14, 15, 17], restricting our attention to the simplest case with a single state preparation and a single, two-outcome measurement. To start, suppose that we have the ability to prepare an (unknown) state ρ, perform an (unknown) two-outcome measurement E, and perform some number n additional (unknown) operations $\{G_0, \dots G_{n-1}\}$. We think about such a system as a box with labeled buttons, as depicted in Figure 1, where each button denotes an operation we can perform. Hence, we have a button for state preparation (b_ρ), measurement (b_E), and buttons for each other operation labeled by elements of the set $\mathcal{B} := \{b_0, \dots, b_{n-1}\}$, where we abbreviate $b_i = b_{G_i}$ for notational convenience. A light on the box turns on or stays off to indicate the outcome of the measurement.

Within this formalism, all experiments we can perform are of the form:

1. Press b_ρ to begin the experiment.
2. Sequentially press zero or more buttons from the set $\{b_0, \dots, b_{n-1}\}$.
3. Press b_E to end the experiment.
4. Record whether the light turned on.

Our goal is to compute the likelihood of observing the light given a particular sequence of buttons. Within a quantum model, we do this by expressing the actions of buttons as *super-operators*, which are linear operators that take density matrices to density matrices. Formally, let $\mathcal{H} = \mathbb{C}^d$ be a Hilbert space of finite dimension d. Then, we denote by $\mathrm{L}(\mathcal{H})$ the space of linear functions $\mathcal{H} \to \mathcal{H}$, and denote by $\mathrm{T}(\mathcal{H})$ the space of linear functions $\mathrm{L}(\mathcal{H}) \to \mathrm{L}(\mathcal{H})$. Since $\mathrm{T}(\mathcal{H})$ is a space of linear functions, elements of $\mathrm{T}(\mathcal{H})$ can be written down as linear operators acting on vectors in $\mathrm{L}(\mathcal{H})$. We denote vectors in $\mathrm{L}(\mathcal{H})$ by "super-kets", e.g. $|\rho\rangle\!\rangle$; covectors for $\mathrm{L}(\mathcal{H})$ are "super-bras" and correspond to measurements, e.g., $\langle\!\langle E|$.

As an example, if $d = 2$, then ρ can be represented as a 2×2 matrix, which we can instead arrange as a 4×1 column vector $|\rho\rangle\!\rangle \in \mathbb{C}^4$. In this case, we can represent elements from $\mathrm{T}(\mathcal{H})$ as 4×4 matrices, which act linearly (by multiplication) on super-kets. [1] We assign to each button b_i a super-operator $G_i \in \mathrm{T}(\mathcal{H})$ (which we represent as a matrix acting on $\mathbb{C}^{d^2}$). If Λ_{G_i} is a quantum channel acting on a density matrix ρ, then G_i is the operator such that $G_i|\rho\rangle\!\rangle = |\Lambda_{G_i}[\rho]\rangle\!\rangle$.

We refer to the set of button presses between applications of SPAM as a *sequence*. We denote the set of possible sequences as $\mathcal{S}$, which contains the empty sequence, as well as every possible combination of button presses. Sequences compose under concatenation. [2] For two sequences $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{S}$:

$$(b_{s_0}, \dots, b_{s_{m-1}}) + (b_{t_0}, \dots, b_{t_{m'-1}}) = (b_{s_0}, \dots, b_{s_{m-1}}, b_{t_0}, \dots, b_{t_{m'-1}}). \quad (1)$$

We will also write $|\boldsymbol{s}|$ to mean the length of $\boldsymbol{s}$, such that $|\boldsymbol{s} + \boldsymbol{t}| = |\boldsymbol{s}| + |\boldsymbol{t}|$.

[1] This is because, if $\mathrm{vec}(A)$ denotes the vectorization of matrix A (by column-stacking), then $\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B)$. As a density matrix ρ evolves by similarity transformation (or a sum over them), $|\rho\rangle\!\rangle$ evolves by matrix multiplication.

[2] The set of experimental sequences $\mathcal{S}$ is a *monoid* under addition; that is, $\mathcal{S}$ is closed under concatenation, and has the empty sequence of buttons $()$ as an additive identity. We note that $\mathcal{S}$ does not contain inverses, since we cannot make a sequence of buttons shorter by pressing more buttons.

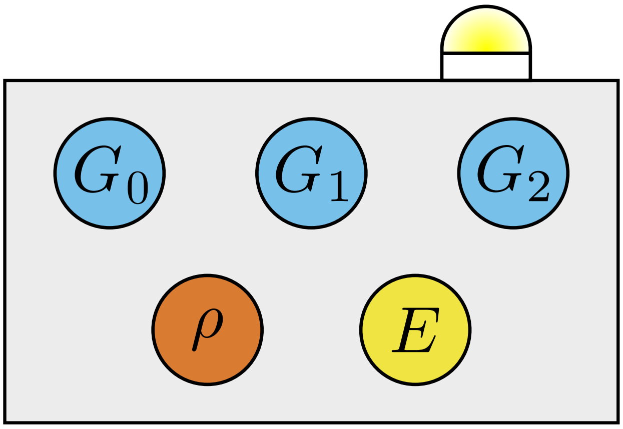

Figure 1: An example of a quantum device modeled as a black box with buttons. Buttons are labeled by the actions they perform, for example prepare state ρ, apply operation G_i, and take measurement E. A light on top of the box turns on or stays off to indicate the result of the measurement.

Using the assignment of super-operators to buttons, we can compute the likelihood of experimental outcomes.

Definition 1. *Let* $\boldsymbol{s} = (b_{s_0}, b_{s_1}, \ldots b_{s_{m-1}})$ *be a sequence of* m *button presses from the buttons on our box. The likelihood of the light turning on after performing* $\mathbf{s}$*, the* sequence probability*, is given by the Born rule:*

$$\Pr(\mathit{light}|(b_\rho, b_{s_0}, \ldots, b_{s_{m-1}}, b_E)) = \langle E|G_{s_{m-1}}G_{s_{m-2}}\cdots G_{s_0}|\rho\rangle. \quad (2)$$

This shows that, were we to learn the explicit form of $|\rho\rangle, \langle E|, \{G_i\}$, we would be able to predict the results of any future experiment. Nevertheless, as we already touched on in the introduction, super-operators suffer from a *gauge* problem, making many numerically distinct sets of super-operators operationally equivalent. (In the language of super-operators, $\{|\rho\rangle, \langle E|, \{G_i\}\}$ is gauge-equivalent to $\{B|\rho\rangle, \langle E|B^{-1}, \{BG_iB^{-1}\}\}$ for any appropriately-sized invertible matrix B.) In the next section, we clarify this notion within the context of linear-inversion GST.

2.2 Linear inversion GST

The simplest GST inference procedure to learn $|\rho\rangle, \langle E|, \{G_i\}$ is linear-inversion GST (LGST). For any GST protocol, one first chooses a set of *fiducial* sequences, $\boldsymbol{f} = \{\boldsymbol{f}_i\}$, which act as a "reference frame" for analysis of the experiments. [3] Fiducial sequences are typically short sequences of button presses, and as a set they must be informationally complete (which will be formally defined below, with a consequence being that the set of fiducials has at least d^2 elements).

We use the set of fiducial sequences to construct the following scalar quantities:

$$\begin{aligned} \tilde{E}_i &= \langle E|F_i|\rho\rangle, \\ \tilde{F}_{ij} &= \langle E|F_iF_j|\rho\rangle, \\ \tilde{G}^{(k)}_{ij} &= \langle E|F_iG_kF_j|\rho\rangle, \end{aligned} \quad (3)$$

where F_i is the super-operator obtained by multiplying together the super-operators for the constituent buttons in a fiducial sequence $\boldsymbol{f}_i$. In principle, the entries of these matrices are the probabilities of the light turning on for the given experiments. Hence, by repeating the experiments, we approximate these probabilities via the empirically observed frequencies. Defining $A = \sum_j |j\rangle\langle E|F_j$ and $B = \sum_j F_j|\rho\rangle\langle j|$, where the $|j\rangle$ are basis states of the space $\mathcal{H} \otimes \mathcal{H}$, we can recover the desired $|\rho\rangle, \langle E|$, and $\{G_k\}$ according to:

$$\begin{aligned} |\rho\rangle &= B\tilde{F}^{-1}\tilde{E} \\ \langle E| &= \tilde{E}^T B^{-1} \\ G_k &= B\tilde{F}^{-1}\tilde{G}^{(k)}B^{-1} \end{aligned} \quad (4)$$

We further note that, by definition, $\tilde{F} = AB$. We require that $\tilde{F}$ has rank of at least d^2, where d is the dimension of the qubit Hilbert space. If $\dim(\tilde{F}) > d^2$, then the pseudo-inverse is used instead of the normal inverse. This rank criterion ($\mathrm{rank}(\tilde{F}) = d^2$) is what provides our definition of informational completeness in this context. It provides a check of the choice of fiducials, which can be useful if good initial guesses of the gates are not known. Such a set of fiducials can even be chosen 'on the fly' by performing experiments until one can construct an invertible $\tilde{F}$.

2.3 Gauge and the operational representation

In the above section, one might be troubled that we did not actually recover the literal values G_k, ρ, and E. Rather, they are now complicated by the presence of the gauge transformation B - the "true" super-operators G_k, ρ, E are all gauge-dependent quantities. However, the gauge B itself is not accessible experimentally. This is because the observed sequence probabilities are totally independent of gauge:

$$\mathrm{Tr}\left(|\rho\rangle\langle E|G_{s_{m-1}}\cdots G_{s_0}\right) = \mathrm{Tr}\left(B^{-1}|\rho\rangle\langle E|BB^{-1}G_{s_{m-1}}B\cdots B^{-1}G_{s_0}B\right) \quad (5)$$

More formally, let us begin by making a mapping between button sequences and super-operators. We assign an element of $\mathrm{T}(\mathcal{H})$, the space of linear operators on $\mathcal{H}$, to each sequence $\boldsymbol{s} \in \mathcal{S}$ using a mapping

[3] It is sometimes necessary to pick distinct *preparation* and *measurement* fiducial sets; our results extend to those situations as well.

$\Phi : \mathcal{S} \to \mathrm{T}(\mathcal{H})$,

$$\Phi((b_{s_0}, \dots, b_{s_{m-1}})) = G_{\boldsymbol{s}}. \tag{6}$$

In general, the mapping Φ between button sequences and channels can arbitrary, especially in the presence of non-Markovian errors, but in this work we will consider the special case in which Φ is a homomorphism between the monoids $\mathcal{S}$ and $\mathrm{T}(\mathcal{H})$ [4],

$$\Phi((b_{s_0}, \dots, b_{s_{m-1}})) := G_{s_{m-1}} \cdots G_{s_0}. \tag{7}$$

This mapping is not unique, but can be specified by the outputs of Φ for each single-button sequence, $\Phi((b_0)) = G_{b_0}$, $\Phi((b_1)) = G_{b_1}$, and so forth. Considering this special case, we can think of SPAM as a special button b_{SPAM} such that

$$\Phi((b_{\text{SPAM}})) = |\rho\rangle\langle E|. \tag{8}$$

Making this identification, we can then use Φ to recover the probabilities in (2) by taking the trace of $\Phi(\boldsymbol{s})$ for each sequence $\boldsymbol{s} \in \mathcal{S}$, so long as we adopt the convention that $\boldsymbol{s}$ begins with b_{SPAM},

$$\Pr(\text{light}|\boldsymbol{s}; \Phi) = \mathrm{Tr}(\Phi(\boldsymbol{s})). \tag{9}$$

The problem of inferring the properties of our box is equivalent to identifying which Φ maps from button sequences to super-operators in a manner that correctly predicts experimental outcomes according to (2). Following this motivation, we define that two mappings $\Phi, \Phi' : \mathcal{S} \to \mathrm{T}(\mathcal{H})$ are *gauge-equivalent* ($\Phi \sim \Phi'$) if and only if they yield the same sequence probabilities for all elements of $\mathcal{S}$. The term "gauge" used to describe the equivalence class $\sim$ is motivated by the observation that

$$\begin{aligned} &\Phi \sim \Phi' \text{ if and only if there exists } B \in \mathrm{GL}(\mathbb{C}^{d^2}) \\ &\text{such that for all } \boldsymbol{s} \in \mathcal{S} : \Phi(\boldsymbol{s}) = B\Phi'(\boldsymbol{s})B^{-1}. \end{aligned} \tag{10}$$

We say that the equivalence class $[\Phi] := \{\Phi' \in \mathrm{Hom}(\mathcal{S}, \mathrm{T}(\mathcal{H}))$ such that $\Phi' \sim \Phi\}$ of gauge-equivalent Φ is the *gauge orbit*. It is easy to identify one such Φ (just choose any invertible matrix of appropriate dimension), but it is expensive to compute an entire equivalence class of distinct ones.

Choosing a gauge to represent a gate set is typically accomplished through nonlinear optimization, in which a gauge is sought that transforms the estimated gate set to be as close as possible (by some metric) to an ideal "target" gate set. This allows for computation of gauge-variant metrics between the estimate and the target (e.g., diamond distance, fidelity). In practice, these procedures can work reasonably well [17], but they scale inefficiently, are not guaranteed to be numerically stable, and are not guaranteed to not get stuck in a local extremum. Thus, as a practical matter, we would like to identify a set of parameters that is necessary and sufficient to identify gauge orbits without having to actually perform a gauge optimization. That is, we seek to parameterize and perform inference on the set of gauge orbits directly:

$$\mathrm{G}(\mathcal{B}, \mathcal{H}) := \mathrm{Hom}(\mathcal{S}, \mathrm{T}(\mathcal{H}))/\!\sim\; = \{[\Phi]\}, \tag{11}$$

where $A/\!\sim$ is the factor set of A defined by the relation $\sim$ as the set of equivalence classes $A/\!\sim\; := \{[a] : a \in A\}$.

When it is clear from context which button set and Hilbert space are used to define our box, we will omit them for brevity, writing that $\mathrm{G} = \mathrm{G}(\mathcal{B}, \mathcal{H})$. We say that each member of $\mathrm{G}(\mathcal{B}, \mathcal{H})$ is a *gate set*, such that identifying which member of $\mathrm{G}(\mathcal{B}, \mathcal{H})$ was used to generate a data record is *gate set tomography*. When it is clear from context, we will also refer to sets of super-operators $\mathcal{G} = \{G_0, \dots, G_{k-1}, |\rho\rangle, \langle E|\}$ as gate sets, with the implicit understanding that we are interested in the gauge orbit $[\mathcal{G}]$ (equivalence class under $\sim$) of $\mathcal{G}$.

We call any such representation of $\mathrm{G}(\mathcal{B}, \mathcal{H})$ *operational*, since it is a complete description of all operational experiments that we can perform on our box, under the promise that the box is described by some model over $\mathcal{H}$. In fact, we have already seen an especially convenient operational representation: $\tilde{E}, \tilde{F}, \{\tilde{G}^{(k)}\}$. They are a set of gauge-independent values (as they are directly experimentally observable), and are unique to a particular gate set for a given choice of fiducials. They also yield the same measurement probabilities as their gauge-dependent counterparts. To see this, consider some sequence of button presses $(b_\rho, b_{s_0}, \dots b_{s_{m-1}}, b_E)$. The sequence probability is:

$$\begin{aligned} &\Pr(\text{light}|(b_\rho, b_{s_0}, \dots, b_{s_{m-1}}, b_E)) \\ &= \langle E|G_{s_{m-1}} \cdots G_{s_0}|\rho\rangle \\ &= \mathrm{Tr}\left(|\rho\rangle\langle E|G_{s_{m-1}} \cdots G_{s_0}\right) \\ &= \mathrm{Tr}\left(B^{-1}|\rho\rangle\langle E|BB^{-1}G_{s_{m-1}}B \cdots B^{-1}G_{s_0}B\right) \\ &= \mathrm{Tr}\left(\tilde{F}^{-1}\tilde{E}\tilde{E}^{\mathrm{T}}\tilde{F}^{-1}\tilde{G}^{(s_{m-1})} \cdots \tilde{F}^{-1}\tilde{G}^{(s_0)}\right). \end{aligned}$$

This leads to the remarkable fact that *when we learn* $\tilde{E}, \tilde{F}, \{\tilde{G}^{(k)}\}$, *we can predict the outcome of any future experiments.* Note that this statement is distinct from performing LGST: we can use $\tilde{E}, \tilde{F}, \{\tilde{G}^{(k)}\}$ as our underlying model, while updating it via more sophisticated experiments.

[4] Note that $\mathrm{T}(\mathcal{H})$ is monoid under multiplication rather than concatenation. In general, $\mathrm{T}(\mathcal{H})$ fails to be a group, as we cannot invert general quantum operations due to decoherence. (Decohering channels are invertible, as long as they are full rank, but these inverses are not completely positive, and are therefore unphysical.) We can then view Φ as a homomorphism from button sequences to super-operators, since $\Phi(s + t) = \Phi(t)\Phi(s)$ and $\Phi(()) = \mathbb{1}$. Since we have listed sequences left-to-right rather than right-to-left, Φ is formally a homomorphism from button sequences to the opposite monoid of super-operators, defined by the opposite product $A \cdot^{\mathrm{op}} B := BA$, but we ignore this detail as a notational convenience.

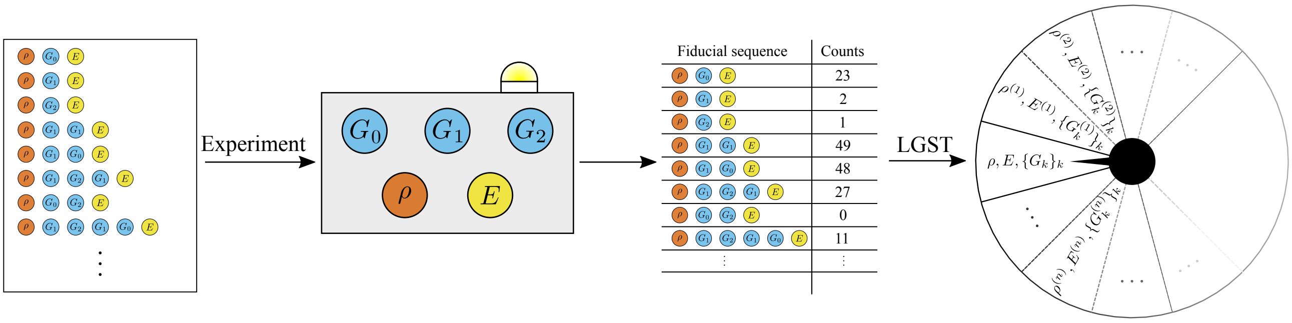
Experiment
G_0
G_1
G_2
ρ
E
Fiducial sequence
Counts
23
2
1
49
48
27
0
11
LGST
$\rho, E, \{G_k\}_k$
$\rho^{(1)}, E^{(1)}, \{G_k^{(1)}\}_k$
$\rho^{(2)}, E^{(2)}, \{G_k^{(2)}\}_k$
$\rho^{(n)}, E^{(n)}, \{G_k^{(n)}\}_k$

Figure 2: Pipeline for linear inversion gate set tomography (LGST). A set of fiducial sequences is chosen; we perform the specified experiments and record how many times the light turned on. Following the linear inversion step in (4), we can reconstruct a copy of the super-operators for each button. However, the results we obtain will be expressed in an unknown gauge which is one of infinitely many in the gauge orbit.

3 Implementation

Having thus established that learning the operational representation of a gate set allows us to predict its behavior, we are left with the question of *how* to learn operational representations from data records. In this section, we describe our implementation of operational quantum tomography, based on Bayesian inference. In particular, we implement the inference numerically using the particle filter, or sequential Monte Carlo (SMC) approach, a standard technique for carrying out Bayesian inference computationally [33].

3.1 Bayesian inference: obtaining posteriors from evidence

As applied to quantum information, Bayesian inference is a formalism for describing our knowledge about a quantum system given classical data observed from it. In particular, Bayesian inference represents our state of knowledge at any given point in a characterization protocol by a distribution of the form $\Pr(\text{hypothesis}|\text{data})$, where "hypothesis" describes some hypothesis that we can use to predict the future behavior of our quantum system, and "data" is the set of observations made of that system.

In the special case that data $= \{\}$ (that is, before we have made any observations), we write our state of knowledge as $\Pr(\text{hypothesis})$, also known as our *prior distribution*. For example, in traditional Ramsey interferometry, our hypothesis might consist of the assumption that the system evolves under a Hamiltonian of the form $H = \omega\sigma_z/2$ for some ω. We may assign a prior distribution over ω such as

$$\Pr(\omega) = \begin{cases} 1/\omega_{\text{max}} & \omega \in [0, \omega_{\text{max}}] \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

representing that we are equally willing to believe that ω has any value in the interval $[0, \omega_{\text{max}}]$.

Since distributions of the form $\Pr(\text{hypothesis}|\text{data})$ represent our state of knowledge at any point during an experimental procedure, equipped with such a distribution, we can answer questions such as "what is the best hypothesis to report given what we have learned from our quantum system?". Returning to the Ramsey example, we may want to report an estimate $\hat{\omega}$ such that the the squared error $(\hat{\omega} - \omega)^2$ is minimized on average. As summarized in Appendix A, this is achieved by reporting the Bayesian mean estimate $\hat{\omega}_{\text{BME}} := \mathbb{E}_\omega[\omega|\text{data}] = \int \omega \Pr(\omega|\text{data}) \mathrm{d}\omega$.

We are thus left with the problem of finding our state of knowledge at some point in an experimental procedure given our most recent observation, and given our previous state of knowledge; that is, of how to update our state of knowledge to reflect new information. To do so, we rely on Bayes' rule, which states that

$$\Pr(\text{hypothesis}|\text{data}) \sim \\ \Pr(\text{data}|\text{hypothesis}) \times \Pr(\text{hypothesis}), \tag{13}$$

where $\Pr(\text{hypothesis})$ is our prior distribution, and $\sim$ indicates equality up to renormalization. Intuitively, this rule tells us that a hypothesis is reweighted according to how plausible it is for a given observation to arise given that hypothesis. To perform this update, we must simulate $\Pr(\text{data}|\text{hypothesis})$, known as the *likelihood function* for our quantum system. Put differently, we can only learn properties of a system whose effects can be simulated. We cannot learn about a parameter that has no effect on the outcomes of system, or whose effects we cannot simulate.

It is for this reason that, in the rest of the paper, we take our hypothesis to be the operational representation of some quantum system. In particular, the operational representation is a minimal set of parameters required to simulate the behavior of that system, such that any parameter beyond the operational representation cannot have any effect on our predictions. For example, we can never learn gauge parameters from experimental observations, as they have no effect on the likelihood function for any measurement that we could perform [5].

3.2 Numerical approach: sequential Monte Carlo

So far, we have regarded Bayesian inference in the abstract, without reference to or concern for how one might implement an inference procedure in practice. A practitioner interested in using Bayes' rule will find it difficult to work with (13) directly, as the normalization suppressed by the use of $\sim$ notation converges exponentially quickly to 0 with the amount of data considered, exacerbating numerical precision issues. Moreover, any choice of discretization informed by the prior is not likely to be terribly useful as the posterior shrinks in width.

In lieu of these considerations, a number of different computational algorithms have been developed that offer a Bayesian practitioner a range of different options. For instance, rejection sampling techniques such as the Metropolis–Hastings algorithm [34], as well as more sophisticated modern algorithms such as Hamiltonian Monte Carlo [35] and NUTS [36], allow for obtaining samples from a posterior distribution with reasonable computational effort. These algorithms have been used in quantum information to solve otherwise intractable problems such as the estimation of randomized benchmarking parameters [37].

For application to online experimental protocols, however, it is often useful to adopt an algorithm that works in a *streaming* fashion. This allows for samples from a posterior distribution to be drawn at any point in an experimental procedure, such that adaptive decisions such as stopping criteria or experiment design can be made easily. Critical to realizing this capability is that the cost of an algorithm can depend only approximately linearly on the amount of data taken. This restriction motivates the use of filtering algorithms, which update an approximation of a prior given incoming data to yield a new approximation of the resulting posterior. The Kalman filter, for example, is a Bayesian filter for the special case in which the prior and posterior are both normal, and in which the likelihood is a linear model perturbed by normally distributed noise [38].

In this paper, we adopt the *particle filter* [33], also known as the sequential Monte Carlo approximation. Particle filters are applicable to a very broad range of likelihood functions, and give rich diagnostic data to assist in understanding their execution. The QInfer library [39] provides a useful implementation of particle filters for quantum information applications, and this library is used throughout the rest of the work.

Particle filters work by representing the distribution over some random variable $\boldsymbol{x}$ as a weighted sum of δ-distributions at each step,

$$\Pr(\boldsymbol{x}) \approx \sum_i w_i \delta(\boldsymbol{x} - \boldsymbol{x}_i), \tag{14}$$

where $\{w_i\}$ are non-negative real numbers summing to 1, and where $\{\boldsymbol{x}_i\}$ are different hypotheses about $\boldsymbol{x}$. Each hypothesis $\boldsymbol{x}_i$ is called a *particle*, and is said to have a corresponding weight w_i. Numerical stability is achieved by periodically moving each particle to concentrate discretization on regions of high posterior density [40]. Examples of this in operation can be seen in videos at https://youtu.be/aUkBa1zMKv4 and https://youtu.be/4EiD8JcCSlQ.

3.3 Setting priors over the operational representation

Within a Bayesian framework, we begin with a statement about our beliefs before starting an experiment. We write this down formally as a *prior distribution*, which gives us a mathematical description of our prior knowledge. In absence of any data from a particular experimental run, a prior distribution π assigns a probability $\pi(\boldsymbol{x})$ to each object of interest $\boldsymbol{x}$ (*e.g.*, the elements of the operational representation).

In experimental QCVV, we typically express our beliefs in terms of gauge-dependent formalisms (*e.g.*, super-operators). Here, we need to translate these prior distributions into a prior over the operational representation, which is gauge-independent. Fortunately, we can easily sample from the prior distribution over the operational representation induced by a distribution over a gauge-dependent representation. Upon choosing a set of fiducial sequences, we proceed to:

1. State the prior over some gauge-dependent representation (*e.g.*, parameters in super-operators).
2. Draw a sample from the gauge-dependent prior.
3. Convert the gauge-dependent sample to the operational representation by applying LGST.
4. Return this as the sample from the gauge-independent prior.

As a concrete example, suppose we intuit that a particular button should perform single-qubit Z-rotation gates. We can write these in a familiar, gauge-dependent way by expressing them as superoperators in some matrix basis, $R_z(\theta)$ (in our implementation, we use the Pauli basis). Now suppose

[5] This argument shows that the use of operational representations can be motivated by appeal to the *likelihood principle*, which informally states that all inference — whether or not carried out using Bayesian reasoning — must depend on a system only through its likelihood function.

that we suspect this button over-rotates about Z by an angle $\delta\theta$ that is somewhere between 0 and $\pi/10$. To generate samples from this prior expressed in the operational representation according to this belief, we first choose samples of $\delta\theta$ uniformly at random from 0 to $\pi/10$. Next, we use these sampled angles to synthesize corresponding channels for each member of the gate set, i.e. $R_z(\theta + \delta\theta)$. A prior distribution in terms of superoperators can be constructed in a similar manner for each button on the box. Together, we use them to compute the frequencies for each element of the operational representation using the linear-inversion step of (4).

3.4 Informational completeness and germ sensitivity

In addition to choosing a prior distribution, we must also choose a set of fiducial sequences to fix a reference frame. Any choice will yield a valid operational representation, in the sense that we can populate an $\tilde{E}, \tilde{F}$, and $\{\tilde{G}^{(k)}\}$ with the outcome frequencies of the experiments. However, an additional requirement of the fiducial sequences is that they must be *informationally complete*. As we will see, the definition of this is dynamic.

Consider for a moment standard quantum state tomography, where we prepare (perfectly) some unknown state, and can execute perfect measurements. In the case of a single qubit, it is well known that measuring σ_x, σ_y, and σ_z is sufficient to fully reconstruct the state [41]. These measurements span the Bloch sphere, and we say that they are informationally complete.

In GST, a similar notion holds. However, we do not know *a priori* how the measurement and operation buttons are oriented relative to an external reference frame. For instance, if someone provides us with a box with buttons labeled σ_x, σ_y, σ_z, we do not know what they actually do. They may be noisy implementations of these operations, they may be completely different operations, or, they may even do nothing at all. Naively using these buttons to execute measurements is therefore not guaranteed to give us something informationally complete, even if the labels suggest they should.

In GST, we can check for informational completeness using the matrix $\tilde{F}$ in the operational representation. Recall that this is constructed using experiments performed by applying pairs of the fiducial sequences. When we initialize the operational representation from the prior over super-operators, we must compute $\tilde{F}^{-1}$. If the fiducial sequences are poorly chosen, $\tilde{F}$ may be ill-conditioned, or even singular.

Definition 2. *A set of fiducial sequences $\boldsymbol{f} \subset \mathcal{S}$ is* informationally complete *for a gate set $\mathcal{G}$ if $\tilde{F} = \sum_{ij} \langle E|F_i F_j|\rho\rangle |i\rangle\langle j|$ has rank of at least d^2, where d is the dimension of the qubit Hilbert space. Here, $F_i = \Phi[f_i]$ and $|\rho\rangle\langle E| = \Phi[\mathrm{SPAM}]$ for some $\Phi \in \mathcal{G}$.*

Note that since we can conjugate by B in Definition 2, we can choose any $\Phi \in \mathcal{G}$ that is convenient for evaluating $\tilde{F}$ — if we can find a Φ such that a set of fiducial sequences is informationally complete, it must also be complete for all $\Phi' \in [\Phi]$.

An issue that arises as a consequence of the choice of fiducials is that it is possible to find values across $\tilde{E}, \tilde{F}, \{\tilde{G}^{(k)}\}$ that are identical. For example, if one of the fiducials is the empty sequence, then $\tilde{E}_0 = \tilde{F}_{00}$, $\tilde{E}_1 = \tilde{F}_{01} = \tilde{F}_{10}$, and so forth. Were we to perform a SMC update step on the full set of matrix entries, the entries that are constrained to be identical will be perturbed in different ways, leading to inconsistent outcomes.

To remedy this, we first perform a preprocessing step that eliminates redundant entries, producing a minimal set of model parameters on which we can perform inference. Mappings are employed to transform the minimal set back to full $\tilde{E}, \tilde{F}, \{\tilde{G}^{(k)}\}$, and *vice versa*, throughout. Learning this minimal set of parameters is then sufficient to characterize the entire system. This trimming procedure also has the benefit of substantially reducing the number of model parameters required, speeding up the inference process.

Beyond fiducial selection, we have considerable freedom in the selection of the particular experiments we perform. The best choice of experiments depends on our particular learning objective. For most of our demonstrations in this work, we fix a total number of experiments, a minimum/maximum sequence length, and then produce sequences of increasing lengths between the bounds. In some implementations of GST, one designs a small collection of short button sequences known as "germs", such that by taking appropriate powers of germs one can amplify coherent errors to gain optimal information as the number of experiments is increased. Such a pattern of germs is called "amplificationally complete" [17], and can reduce the total number of experiments required. We take this approach in our implementation of GST, and take care to identify the experiments we carry out in all of our examples.

3.5 Constraints on gates and gate sets

If we represent inferred channels with superoperators (as is customary in quantum process tomography), the allowed form of the matrices is not arbitrary. Rather, physical constraints such as positivity and trace preservation of density matrices restrict the allowed structure. When generalizing to GST, the problem of identifying when a gate set is valid is complicated by the introduction of the gauge. The elements of a gate set might not, in a particular representation, be CPTP, but are gauge equivalent to a CPTP representation. Performing inference on an operational representation introduces a similar challenge: how do we ensure that an operational representation corresponds to a gate set that makes physical sense?

Analogous to the case in GST, we need a condition that is simultaneously satisfied by all the gates in the gate set. For the operational representation, an obvious first test is to check whether all the entries are in the interval $[0, 1]$. Since entries in the operational representation correspond to sequence probabilities, this is a necessary physical constraint. This is not a sufficient condition, though, as the probabilities for any possible future experiment must be constrained in the same way. This leads us to the following definition:

Definition 3 (Positivity)**.** *An estimate of a gate set* $\hat{\mathcal{G}} = \{\hat{G}_0, \dots, \hat{G}_{n-1}, |\hat{\rho}\rangle\!\rangle, \langle\!\langle \hat{E}|\}$ *is* positive *if for all* $\hat{\mathcal{S}} \in \{\hat{G}_0, \dots, \hat{G}_{n-1}\}^\star$*, where* $\{\cdot\}^\star$ *is the Kleene-closure*[6]*, we have that both* $\langle\!\langle \hat{E}|\hat{\mathcal{S}}|\hat{\rho}\rangle\!\rangle \geq 0$ *and* $\langle\!\langle \mathbb{1} - \hat{E}|\hat{\mathcal{S}}|\hat{\rho}\rangle\!\rangle \geq 0$.

Other than by converting to the operational representation a standard (gauge-variant) representation which is explicitly positive (in some gauge), it is unclear how one can create operational representations that are positive by construction. However, it *is* possible to ensure that inference begins from a point where this is true through our choice of prior distribution. As described in Section 3.3, when we set a prior distribution, we begin with a gauge-dependent prior. When we do this, we express each gate of the gate set in the same gauge that we have chosen. We can then guarantee by construction that each member of the gate set has characteristics such as complete positivity, to ensure they will always produce valid outcome probabilities.

As inference proceeds, however, checking for properties such as complete positivity is practically difficult. This is because to check such properties, one needs to perform a gauge-fixing procedure, which we wish to avoid for aforementioned reasons. Such a procedure is not impractical to do once at the end of an inference procedure, but it is at each update step during Bayesian inference.

One workaround to this is to ensure at the very least all the values in the operational representation are positive. Though this of course doesn't guarantee true positivity, we found in practice that negative values of the likelihood function appear regularly, and these must be handled appropriately in order for the sequential Monte Carlo updates to succeed. As a workaround, we simply clip the output of the likelihood function so that any negative 'likelihoods' are set to 0, and any positive likelihoods greater than 1 are set to 1.

An alternate way to approach model validity is to choose a set of validation experiments. In plausible experimental settings, one has a specific application (and hence gate sequence) in mind. We can then decide if a model is valid for a particular set of gate sequences by checking if it produces a proper likelihood for all the validation experiments, a notion that we call *operational positivity*.

Definition 4 (Operational positivity)**.** *An estimate of a gate set* $\hat{\mathcal{G}} = \{\hat{G}_0, \dots, \hat{G}_{n-1}, |\hat{\rho}\rangle\!\rangle, \langle\!\langle \hat{E}|\}$ *is* operationally positive *on a set* $\hat{\mathcal{S}}_{\text{test}} \subseteq \{\hat{G}_0, \dots, \hat{G}_{n-1}\}^\star$ *if for all* $\hat{\mathcal{U}} \in \hat{\mathcal{S}}_{\text{test}}$ *both* $\langle\!\langle \hat{E}|\hat{\mathcal{U}}|\hat{\rho}\rangle\!\rangle \geq 0$ *and* $\langle\!\langle \mathbb{1} - \hat{E}|\hat{\mathcal{U}}|\hat{\rho}\rangle\!\rangle \geq 0$.

From these definitions, positivity implies operational positivity but the converse need not be true. Operational positivity is a useful concept because it is both easy to test and also of practical relevance when one wishes to check particular applications. In our work, we make extensive use of operational positivity, since whenever the sequential Monte Carlo inference procedure resamples particles, it is necessary to avoid negative predicted probabilities.

4 Prediction Loss

Once we have obtained a posterior distribution $\Pr(\boldsymbol{x}|\text{data})$ over the operational representation $\boldsymbol{x}$ of our gate set from some sequence of experiments, we are typically interested in extracting diagnostic and benchmarking information. To do so in a manner consistent with the gauge, one could consider *gauge fixing* procedures, which consist of optimization problems that pick out particular gauge-dependent representations of a gate set that we can then use to report traditional metrics [17].

For instance, if we intend *a priori* that the b_x, b_y, b_z buttons should be describable by unitary transformations $\mathrm{e}^{-\mathrm{i}\pi\sigma_x/2}$, $\mathrm{e}^{-\mathrm{i}\pi\sigma_y/2}$, and $\mathrm{e}^{-\mathrm{i}\pi\sigma_z/2}$, respectively, we may wish to report gauge-dependent metrics such as the diamond norm by fixing to a gauge that best agrees with this description. By taking the best case over members of the gate set in that gauge we can construct statements such as "there exists a gauge-dependent description of our gates such that with posterior probability at least $(1 - \alpha)$, the agreement between each gate and their action in a particular chosen frame is no worse than ϵ."

Unfortunately, this gauge-fixing procedure can be cumbersome to implement (especially across many hypotheses), is not guaranteed to work (i.e., find the optimum gauge given a target gate set) and is open to multiple interpretations. As an alternative, we instead will score our predictions on a set of experiments of interest. To do this, we recall that a gate

[6] The Kleene closure $S^\star$ of a set $S = \{s_0, s_1, \dots\}$ is given by the set of all finite-length strings over S, $S^\star = \{(), (s_0), (s_1), \dots, (s_0, s_0), (s_0, s_1), \dots, (s_1, s_0), \dots, (s_0, s_0, s_0), \dots\}$.

set $\mathcal{G}$ is sufficient to predict the outcome of *any* hypothetical experiment we may wish to perform within the GST framework by 4. We thus take a data-driven approach to the problem and choose a set of button sequences S_{validate} that we are interested in correctly predicting. Concretely, let $p_{\boldsymbol{s}} := \Pr(\text{light}|\boldsymbol{s})$ for each $\boldsymbol{s} \in S_{\text{validate}}$ be a parameter that we are interested in estimating. If we predict $\hat{p}_{\boldsymbol{s}}$ for $p_{\boldsymbol{s}}$, then we can consider the quadratic loss

$$L_{\boldsymbol{s}}(\hat{p}_{\boldsymbol{s}}, p_{\boldsymbol{s}}) = (\hat{p}_{\boldsymbol{s}} - p_{\boldsymbol{s}})^2 . \qquad (15)$$

We call this a *prediction loss* for the sequence $\boldsymbol{s}$, since it rewards estimators that can accurately predict the outcome of future experiments. The quadratic loss is by no means unique, and there are other suitable choices, such as the Kullback-Liebler divergence.

Since each prediction loss function is Bregman for each $\boldsymbol{s}$, the Bayesian mean estimator (BME), where we average over the prediction made for each gate set in the support of our posterior, is optimal [42] [7]. That is, to minimize loss we choose as our estimator

$$\hat{p}_{\boldsymbol{s}} = \mathbb{E}_{\mathcal{G}}\left[\Pr(\text{light}|\mathcal{G}; \boldsymbol{s})|\text{data}\right] . \qquad (16)$$

Intuitively, we predict the outcome of measuring the sequence $\boldsymbol{s}$ for each hypothesis $\mathcal{G}$, and then take the average.

This gives us much better predictive capability than restricting ourselves to using a single estimated gate set to predict all future experiments. As we validate with longer and longer sequences than those in the training set that we used to obtain our posterior in the first place, our posterior uncertainty in $p_{\boldsymbol{s}}$ will necessarily grow, as can be seen from the method of hyperparameters [44]. This is not reflected if we pick a single gate set, but is immediately included in the Bayesian mean estimator (16), which will tend to hedge towards 1/2 as sequences grow in length.

5 Examples

In this section, we demonstrate the versatility of our framework by applying it to many common QCVV protocols. This includes replicating the results of other state-of-the-art techniques, such as long-sequence gate set tomography [17] and randomized benchmarking. A discussion of applications of OQT to quantum state tomography can be found in Appendix B.

5.1 Ramsey interferometry

Single-qubit operations are often implemented by applying electromagnetic pulses to induce rotations about the Bloch sphere. The basis of such methods is the intrinsic Rabi oscillation frequency of the system, which tells us the likelihood of measuring the qubit in its $|0\rangle$ or $|1\rangle$ state at a given time. Knowledge of the Rabi frequency allows us to adjust the pulse duration in order to obtain exactly the superpositio nwe desire.

Typically, one learns this frequency by means of either Rabi or Ramsey interferometry. In Ramsey interferometry (depicted in Figure 3), a qubit is prepared in state $|0\rangle$ and then a $R_x\left(\frac{\pi}{2}\right)$ pulse is applied. The qubit is left to evolve under the Hamiltonian $H = \frac{\omega t}{2}\sigma_z$ for some time t, after which another $R_x\left(\frac{\pi}{2}\right)$ pulse is applied, followed by measurement in the computational basis. The likelihood of obtaining the measurement outcome $|0\rangle$ is given by

$$\begin{aligned}\Pr(0|\omega; t) &= |\langle 0| R_x\left(\frac{\pi}{2}\right) e^{-i\omega t\sigma_z/2} R_x\left(\frac{\pi}{2}\right)|0\rangle|^2 \\ &= \cos^2\left(\frac{\omega t}{2}\right) \qquad (17)\end{aligned}$$

However, in a given implementation, it is likely that the $R_x(\frac{\pi}{2})$ pulse is not perfect and the resultant state will be slightly over- or under-rotated, yielding an incorrect estimate of ω. Our goal is to learn not only ω, but also the precise rotation angle so that we can compensate for this discrepancy by adjusting the duration of the pulse.

First, we translate Ramsey interferometry into the operational framework language (a box with buttons). In this case, there are four buttons. The first two are for SPAM - the button b_ρ prepares the $|0\rangle$ state, and b_E performs a measurement in the computational basis. The third button b_{R_x} performs $R_x(\frac{\pi}{2})$, and the final button $b_{\delta t}$ waits for a discrete time δt (free evolution). By applying $b_{\delta t}$ a total of n times, we can wait time $t = n \cdot \delta t$.

Next, we choose a prior distribution from which to sample to begin Bayesian inference. For convenience, these priors are summarized in Table 1. As explained in Section 3.3, our prior is defined initially in a gauge-dependent way, which is then used to induce a prior distribution on the gauge-independent operational representation. The initial $R_x\left(\frac{\pi}{2}\right)$ are sampled from a distribution that encompasses over- and under-rotation: we choose rotations of the form $R_x\left(\frac{\pi}{2} \pm \delta\theta\right)$, where $\delta\theta$ is a deviation sampled from a normal distribution with mean 0 and a small variance $\sigma^2 = 10^{-3}$. As δt is meant to indicate evolution around the z axis, we sample from $R_z(\theta)$ with θ chosen uniformly from between 0 and 1.

For both the state preparation and measurement priors, we apply depolarization to the ideal state $|0\rangle$. When acting on a density matrix ρ, depolarization of strength p, $0 \leq p \leq 1$, sends

$$\rho \to (1-p)\rho + \frac{p}{3}\left(X\rho X + Y\rho Y + Z\rho Z\right) . \qquad (18)$$

[7] For a more detailed discussion of the role of Bregman estimators in tomography, we refer the reader to the work of Kueng and Ferrie [43].

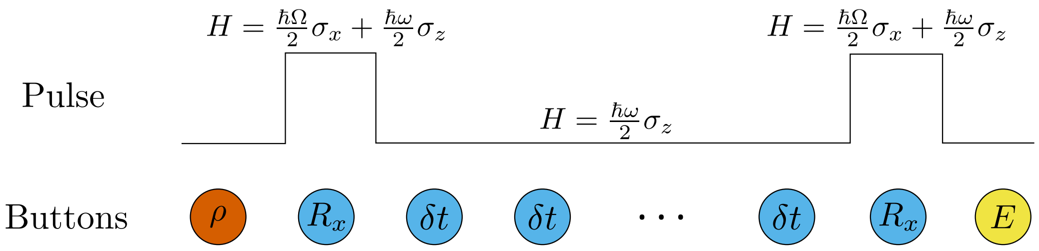
$H = \frac{\hbar\Omega}{2}\sigma_x + \frac{\hbar\omega}{2}\sigma_z$
$H = \frac{\hbar\Omega}{2}\sigma_x + \frac{\hbar\omega}{2}\sigma_z$
Pulse
$H = \frac{\hbar\omega}{2}\sigma_z$
Buttons
ρ
R_x
δt
δt
$\cdots$
δt
R_x
E

Figure 3: Depiction of Ramsey interferometry as a pulse diagram, and as a button sequence in the OQT framework.

Button label	**Prior**	**Example values**
ρ	$1/\sqrt{2}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix}$, depolarized with $p \in \mathcal{U}(0, 0.1)$	$p = 0.038311$
R_x	$R_x\left(\pi/2 + \epsilon\right), \quad \epsilon \in \mathcal{N}(0, 10^{-3})$	$\epsilon = -0.003824$
δt	$R_z(\omega \cdot \delta t), \quad \omega \in \mathcal{U}(0, 1)$	$\omega = 0.346754, \quad \delta t = 1$
E	$1/\sqrt{2}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix}$, depolarized with $p \in \mathcal{U}(0, 0.1)$	$p = 0.023933$
Fiducial seqs.	$\{(\cdot), \quad (R_x), \quad (R_x,\ R_x), \quad (R_x,\ \delta t,\ R_x)\}$	
Training exps.	$(R_x,\ (\delta t)^n,\ R_x)$	$n = 2, \ldots, 49$
Testing exps.	$(R_x,\ (\delta t)^n,\ R_x)$	$n = 50, \ldots, 100$

Table 1: OQT parameter specification for Ramsey interferometry. Example values correspond to those used in the plots and provided example notebook. Button sequences are represented as lists, where the buttons are applied from left to right, and application of SPAM is implicit in the training and testing experiments. Button labels are abbreviated as $b_{R_x} \to R_x$ for notational simplicity.

The associated Bloch vector then transforms according to [41]:

$$(a_x, a_y, a_z) \to ((1-p)a_x, (1-p)a_y, (1-p)a_z). \quad (19)$$

Here all super-operators are expressed in the Pauli basis, where applying depolarization to super-operator G sends

$$G \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1-p \end{pmatrix} G. \quad (20)$$

For the priors, we assume that depolarization occurs with strength p chosen from the uniformly at random between 0 and 0.1, denoted $\mathcal{U}(0, 0.1)$.

The last step is to choose a set of fiducial sequences. We choose $f = \{(\cdot), \ (b_{R_x}), \ (b_{R_x}, b_{R_x}), \ (b_{R_x}, b_{\delta t}, b_{R_x})\}$. These sequences are read from left to right; the first is the empty sequence, and application of SPAM is implicit in all sequences. This choice is not unique, and we picked some that performed well in practice. Using these fiducials to construct an operational representation results in 27 parameters, which is reduced due to duplication from the 52 that are expected from counting the full $\tilde{E}, \tilde{F}$, and $\tilde{G}^k$.

We initialized a SMC cloud with 10000 particles, and performed Bayesian inference over these parameters by feeding in simulated experimental data for sequences of the form $(b_{R_x}, \ (b_{\delta t})^n, \ b_{R_x})$ for $n = 2, \ldots 49$. The 'true' values of the parameters that generated the data were randomly sampled from the prior distribution, with specific parameters given in Table 1. See the supplementary materials for the implementation.

In Figure 4 we plot the likelihood as calculated over the posterior distribution, and compare to the true likelihood (in this case, the set of model parameters that was used to produce the experimental data). We see that our inference has learned this operational representation, and produces comparable likelihoods even out to sequences that are double the length of those we trained with. Using Figure 4, it is possible to fit a curve of the form $\cos^2(\omega t/2)$ and extract an estimate for the value of ω. We obtain $\hat{\omega} = 0.345905$, a roughly 0.6% difference from the true value $\omega = 0.346754$ as noted in Table 1.

Instead, we can judge the quality of the reconstruction by plotting prediction loss, as shown in the right panel of Figure 4. The amount of quadratic loss is small in the absolute sense, and clearly worsens with experiment length, with the peaks increasing quadratically. To build upon Section 4, we include in the supplementary materials a similar plot using the KL divergence.

We can also visually examine the loss by plotting trajectories of different particles sampled from the posterior. Shown in Figure 5 are the trajectories of 50 such particles, with likelihoods computed out to sequences of up to $n = 300$ presses of $b_{\delta t}$. As one might expect, we can see that the trajectories begin

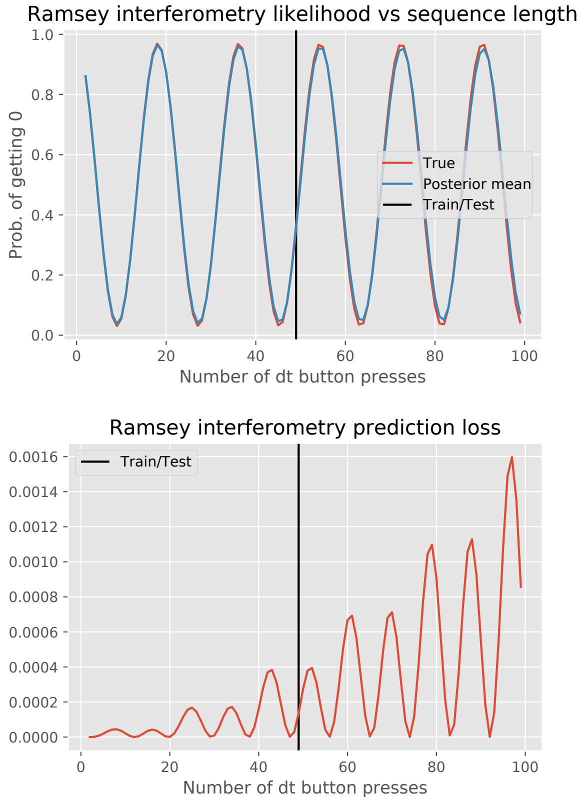

Figure 4: (Top) Likelihood vs sequence length for the true gate set compared to the gate set obtained by taking the mean over the posterior distribution. The mean posterior matches closely up to the testing point, and then begins to diverge. Fitting the curve produced $\hat{\omega} = 0.345905$, which is a roughly 0.6% difference from the true value. (Bottom) The divergence can be quantified using a prediction loss. Shown here is the quadratic loss, $(\hat{p}_s - p_s)^2$. While small, it increases steadily as the sequence length increases.

to ‘spread’ significantly past the $n = 100$ point. The spread can also be quantified and visualized in the manner of Figure 6, in which we have plotted the difference between the likelihoods of all particles of the posterior and the true likelihoods at each sequence length. We can see that the mean deviation from the likelihood increases as the spread in possible values becomes greater at longer sequence length.

While Ramsey interferometry is an arguably simple characterization procedure, it is perhaps the most surprising successful application of OQT we have explored. The same task would not be possible in the traditional GST formalism if one is limited to performing only Ramsey-type experiments, namely two $R_x(\frac{\pi}{2})$ pulses separated by some amount of time. Circuits of that form are not rich enough to generate all the sequences required by GST. While one *can* construct an informationally complete fiducial set using only compositions of $R_x(\frac{\pi}{2})$ and $R_z(\delta t)$ gates, there will always be GST-required circuits that do not follow the Ramsey circuit form. (For example, GST will require at least one circuit that requires *three* applications of $R_x(\frac{\pi}{2})$; such a circuit is not allowed if one is *only* performing Ramsey circuits, all of which have only two applications of $R_x(\frac{\pi}{2})$.) Even though such circuits appear in the operational representation, our prior information allows us to not perform them if we so choose; this highlights the value of being able to incorporate prior information into a characterization protocol. Since the entire prior distribution is created computationally, we can still perform OQT even in cases where we are not able to physically perform a set of experiments that corresponds to every sequence in the operational representation.

5.2 Long-sequence gate set tomography

We next compare OQT to long-sequence GST, where carefully designed sequences are used to self-consistently fit both SPAM and an unknown gate set [17]. Long-sequence GST uses the linear-inversion step of LGST as a starting point, and then proceeds with a longer maximum-likelihood estimation over experiments with progressively longer sequences of gates. Once the procedure finishes, a final gauge fixing is often used to compare the resulting super-operators to expected super-operators.

In [17], long-sequence GST was performed on experimental data from a trapped-ion qubit on which we could perform three operations: G_i, $G_x = R_x\left(\frac{\pi}{2}\right)$, and $G_y = R_y\left(\frac{\pi}{2}\right)$. Thus including SPAM, our box has five buttons.

The linear inversion step in [17] was originally performed using 6 fiducials. However, choosing the same 6 fiducials here results in a 6×6 $\tilde{F}$ that has rank 4. The reason to include those extra fiducials is to increase stability, since LGST then represents an overdetermined system of linear equations. In OQT, we can still include these extra experiments in our analysis, but since the fiducials are used directly to define our model parameters, we need to pick a subset of of size 4 (we choose $\boldsymbol{f} = \{(\cdot),\ (b_{G_x}),\ (b_{G_y}),\ (b_{G_x}, b_{G_x})\}$).

We perform OQT using the set of experiments included in the supplementary material of [17]. These experiments have the form $(\boldsymbol{f}_i, (\boldsymbol{g}_k)^L, \boldsymbol{f}_j)$, where $\boldsymbol{g}_k$ are ‘germ’ sequences that are specified in Table 2. The particular germs were chosen in [17] because they are *amplificationally complete*. From these experiments, we do not include those of the form $((b_{G_x})^n)$, $((b_{G_y})^n)$, and $((b_{G_i})^n)$ for $n = 1, 2, 4, \ldots 8192$ in our training data set – these are kept as a testing set.

The choice of prior plays a particularly important role here, due to the inherently noisy nature of a physical system. We choose a very general prior, based on convex combinations of the ideal super-operators with ones chosen uniformly at random. For both ρ and E, we take a combination of the form

$$\rho' = (1 - \epsilon)\rho + \epsilon\sigma, \quad \epsilon = 10^{-4}, \quad \sigma \in \mathrm{GinibDM}(2). \tag{21}$$

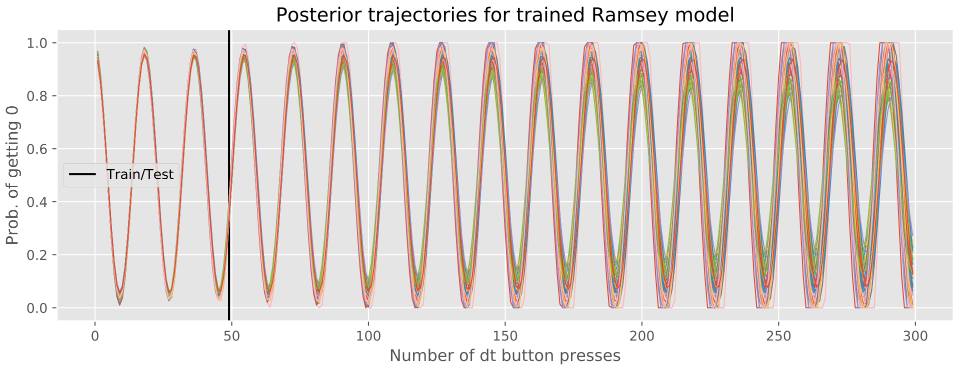

Figure 5: The trajectories of 50 particles sampled from the posterior operational representation for Ramsey interferometry. We see an increased spread at higher sequence lengths, which is highlighted later in Figure 6.

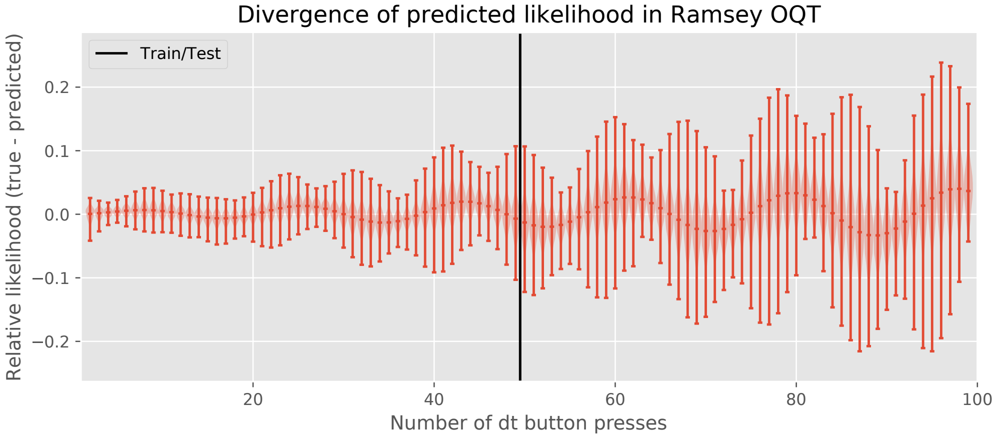

Figure 6: As the sequence length increases, the spread of likelihoods increases as well. Shown in this violin plot is the distribution of 'relative likelihood' for particles in the posterior distribution at each sequence length, i.e. $(\hat{p}_s - p_s)$.

GinibDM(2) denotes the Ginibre distribution, the uniform distribution over single-qubit density matrices. Such states are sampled by computing

$$\sigma = \frac{XX^\dagger}{\mathrm{Tr}(XX^\dagger)}, \quad X_{ij} = a + bi, \;\; a, b \in \mathcal{N}(0,1), \quad (22)$$

where here X is a 2×2 matrix. We take a similar approach for G_i, G_x, and G_y by adding Ginibre noise to the ideal super-operators:

$$G' = (1-\epsilon)G + \epsilon\Lambda, \quad \epsilon = 10^{-4}, \quad \Lambda \in \mathrm{BCSZ}(2). \quad (23)$$

Here, Λ is a super-operator chosen from the uniform distribution over CPTP super-operators, known as the BCSZ distribution [45], denoted by BCSZ(2).

The choice of ϵ was informed by a combination of the experimental data and a grid search. Observing Figure 1c in [17], we note that likelihoods in the (ideally) definite-outcome testing experiments start to significantly decay at around 10^4 gates, hence we intuit that ϵ should be around 10^{-4}. This was later confirmed using a grid search. We ran OQT using a cloud of 10,000 particles for 192 different combinations of ϵ's. We set ϵ the same for G_i, G_x, G_y at $\{10^{-m},\; 2\cdot 10^{-m},\; 4\cdot 10^{-m},\; 8\cdot 10^{-m}\}$ for $m = 5,4,3$. For SPAM, we also choose ϵ the same for ρ and E, and explore the range $\{10^{-m},\; 2\cdot 10^{-m},\; 4\cdot 10^{-m},\; 8\cdot 10^{-m}\}$ for $m = 5,4,3,2$.

The quality of each pairing of ϵ was determined by (a) whether or not the SMC updater succeeded without all particle weights going to zero, and (b) the sum of the total variation distance over the testing experiments. For a given sequence s, let $p(s,\mathcal{R})$ and $p(s,\mathcal{E})$ be the experimental probabilities for a given reconstruction $\mathcal{R}$ and the experimental data $\mathcal{E}$. The total variation distance (TVD) between the two probability distributions is:

$$\mathrm{TVD}(s,\mathcal{R},\mathcal{E}) = |p(s,\mathcal{R}) - p(s,\mathcal{E})|. \quad (24)$$

Bayesian inference ran to completion[8] for ϵ of the gates in the range $4\cdot 10^{-5}$ up to $2\cdot 10^{-4}$. For these

[8] We note that the larger values of ϵ for which inference did not complete in this case may still yield results if the number of particles is increased, given that the noisy super-operators obtained with smaller ϵ will still be in the support of the prior. This highlights the trade-offs one can explore between time, computational resources, and the strength of our assumptions about the buttons.

Table 2: OQT parameters for long-sequence GST on trapped-ion data. Button labels are abbreviated $b_{G_x} \to G_x$ for simplicity here when specifying button sequences. All priors involve adding random noise to the original super-operators using the Ginibre distribution over qubits for ρ and E (denoted here by GinibDM(2)), and the Ginibre distribution for the super-operators (denoted by BCSZ(2)). The set of training experiments is the same as in Blume-Kohout *et al.* [17], however as denoted below we have removed a subset of these for testing.

Button label	**Prior**	**Example values**
ρ	$(1-\epsilon)\left\|0\right\rangle\left\langle 0\right\| + \epsilon\sigma, \quad \sigma \in \mathrm{GinibDM}(2)$	$\epsilon = 10^{-4}$
G_x	$(1-\epsilon)R_x(\frac{\pi}{2}) + \epsilon\Lambda, \quad \Lambda \in \mathrm{BCSZ}(2)$	$\epsilon = 10^{-4}$
G_y	$(1-\epsilon)R_y(\frac{\pi}{2}) + \epsilon\Lambda, \quad \Lambda \in \mathrm{BCSZ}(2)$	$\epsilon = 10^{-4}$
G_i	$(1-\epsilon)\mathbb{1} + \epsilon\Lambda, \quad \Lambda \in \mathrm{BCSZ}(2)$	$\epsilon = 10^{-4}$
E	$(1-\epsilon)\left\|0\right\rangle\left\langle 0\right\| + \epsilon\sigma, \quad \sigma \in \mathrm{GinibDM}(2)$	$\epsilon = 10^{-4}$
Fiducial seqs.	$\{(\cdot),\ (G_x),\ (G_y),\ (G_x, G_x)\}$	
Training exps.	$(\boldsymbol{f}_i, (\boldsymbol{g}_k)^{L_k}, \boldsymbol{f}_j)$ for all fiducials and 11 germs $\boldsymbol{g}_k \in \{(G_x), (G_y), (G_i, G_x, G_y), (G_x, G_y, G_i), (G_x, G_i, G_y), (G_x, G_i, G_i), (G_y, G_i, G_i), (G_x, G_x, G_i, G_y), (G_x, G_y, G_y, G_i), (G_x, G_x, G_y, G_x, G_y, G_y)\}$ (unique sequences only, with testing sequences removed)	$L_k = \left\{\left\lfloor \frac{2^m}{\|g_k\|} \right\rfloor\right\}, m = 1, ..., 13$
Testing exps.	$((G_x)^n)$, $((G_y)^n)$, and $((G_i)^n)$	$n = 1, 2, 4, ..., 8192$

values, inference was successful over essentially the full range of SPAM values. However the sum of total variation distances was notably lower for gate $\epsilon = 10^{-4}$, and SPAM ϵ between 10^{-5} and 10^{-4}, reaching a minimum of 10^{-4} during one full sweep of the grid search.

Results for OQT run with the parameters of Table 2 are plotted in Figure 7. The left column of plots compares the likelihoods predicted for the test sequences from the OQT posterior distribution to the likelihoods of the 'perfect' gate, the experimental counts, and the gate set reconstructed in [17] using the pyGSTi software package. We see that OQT produces results that are competitive with its contemporaries without the need to perform MLE. This is quantified in the right column that plots the variation distance for the same set of experiments. The total TVD for the OQT posterior is 0.724, and that of the pyGSTi reconstruction is 0.961.

5.3 Randomized benchmarking

Like GST, OQT equips us to make predictions about the outcome of any future experimental sequences. Hence, as has been done before using GST [17], we can use OQT to perform randomized benchmarking (RB). To do so, we perform OQT to learn the generators of the Clifford group. Then, using samples taken from the obtained posterior, we will apply RB type sequences and compute the survival probability.

5.3.1 Background for randomized benchmarking

RB makes use of random elements of the Clifford group, which for one qubit is constructed using two generators, $\mathcal{C} = \langle H, S \rangle$, where

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \text{and} \quad S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}. \tag{25}$$

Up to a phase, $\mathcal{C}$ contains 24 elements.

A traditional RB experiment seeks to characterize the errors present in our Clifford gates. We begin by preparing a known state ρ_ψ, and then apply a randomly chosen sequence of Clifford elements. This is followed by applying the element that is the inverse of the group element formed by the sequence (not just performing the sequence backwards). We then measure our system using the measurement operator E_ψ corresponding to our initial state.

If there are no errors in the Clifford gates, the action of the sequence and its inverse would cancel, leaving the state exactly as we found it. When there are errors, however, we can compute what is termed the *survival probability* of the original state. As the sequences increase in length, the survival probability decays, as errors accumulate. Typically, one plots a "decay curve" of the form

$$P(m) = (A - B)p^m + B, \tag{26}$$

where m is the sequence length, and where $P(m)$ is the mean survival probability over all sequences of length m. That is, we define that

$$P(m) := \mathbb{E}_{\boldsymbol{s} \in \text{s. t.} |\boldsymbol{s}| = m} \left[\Pr(0|[\Phi]; \boldsymbol{s})\right]. \tag{27}$$

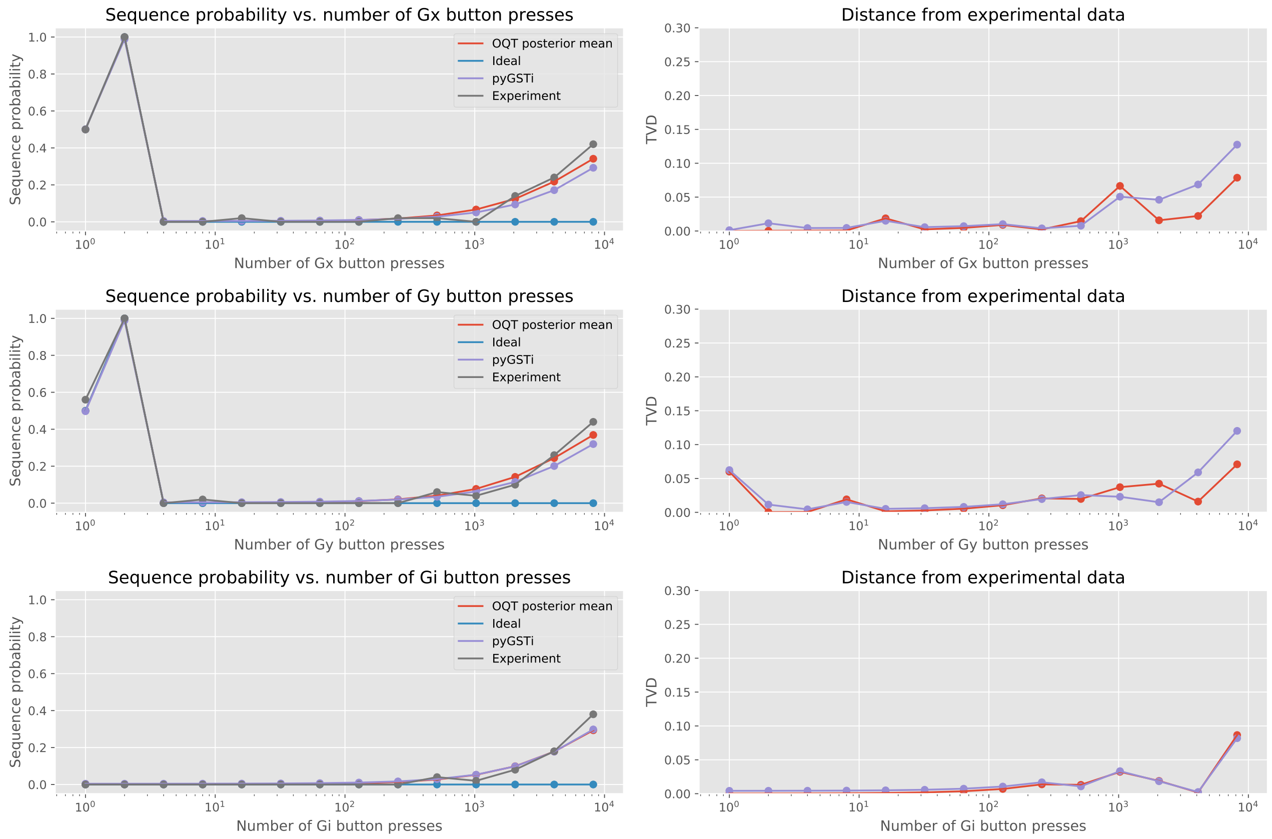

Figure 7: (Left) Comparison of OQT posterior likelihoods obtained using parameters in Table 2 to ideal likelihoods, pyGSTi reconstruction, and experimental data for the trapped-ion data of [17]. The testing experiments consist of exponentially longer sequences of repeated button presses, G_x^k and G_i^k. (Right) Total variation distance of pyGSTi and OQT reconstructions for the same gate sequences. Here we see that the OQT results vary from those of pyGSTi for G_x and G_y, but give comparable results for G_i. The total TVD for OQT across all testing experiments is lower, at 0.724, while that of pyGSTi is 0.961.

We note that since probabilities are *not* directly observable, and can only be estimated, caution must be taken when estimating $P(m)$ or interpreting estimates obtained in an *ad hoc* fashion.

Keeping this caution in mind, the form (26) for the expectation value of the survival probability over sequences of length m was derived analytically by Magesan *et al.* [46], where it was noted that the parameter p contains information about the average fidelity of our Clifford elements. In particular,

$$p = \frac{dF_{\text{ave}}(\Lambda) - 1}{d-1}, \tag{28}$$

where d is the dimension of the Hilbert space under consideration ($d = 2$ for a single qubit RB experiment), where $F_{\text{ave}}(\Lambda)$ is the *average gate fidelity* of the channel Λ, and where Λ is the average error in implementing each member of the Clifford group. In particular, Λ takes on the gauge-dependent form

$$\Lambda = \mathbb{E}_{U\sim\mathcal{C}}[(U^\dagger \bullet)\Lambda_U], \tag{29}$$

where $(U^\dagger \bullet)$ is the ideal action of $U^\dagger$, and Λ_U is a super-operator representing the actual implementation of U.

Despite the large literature on RB [46–60], both the experimental implementation and statistical interpretation are challenging. Since RB is frequently used to assess suitability for quantum error correction applications, this is troubling. Since the technique that we describe here *indirectly* performs RB through *ex post facto* simulation, we are less vulnerable to some of these challenges. More details on this can be found in Appendix C.

5.3.2 Performing RB using OQT

To perform RB using OQT, we need a box with 4 buttons: ρ_ψ, E_ψ, b_H, and b_S. As RB is robust to SPAM errors, we assume for simplicity that SPAM is perfect. That is, we focus on learning H and S. Our first step is to choose an appropriate prior over H and S: we pick one that represents our belief that errors in each generator are due to both systematic over-rotations and Ginibre noise. To apply over-rotation to the Hadamard, we begin with its super-operator representation $G_H = H \otimes H$. Mathematically, this can also be written as $H \otimes H = e^{\frac{i\pi}{2}(H\otimes\mathbb{1} - \mathbb{1}\otimes H)}$, which we recognize as just evolution for the time $\pi/2$ under

Table 3: OQT parameters for randomized benchmarking. Button labels are abbreviated $b_H \to H$ and $b_S \to S$ for simplicity when specifying button sequences.

Button label	**Prior**	**Example values**
ρ_ψ	$1/\sqrt{2}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix}$	Perfect
H	$(1-\epsilon)G_H(\delta\theta_H)+\epsilon\Lambda_H,\quad \delta\theta_H \in \mathcal{N}(0,0.0015),\quad \Lambda_H \in \mathrm{BCSZ}(2)$	Eqs. (30),(34); $\epsilon = 10^{-3}$
S	$(1-\epsilon)R_z(\pi/2+\delta\theta_S)+\epsilon\Lambda_S,\quad \delta\theta_S \in \mathcal{N}(0,0.0015),\quad \Lambda_S \in \mathrm{BCSZ}(2)$	Eq. (34); $\epsilon = 10^{-3}$
E_ψ	$1/\sqrt{2}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix}$	Perfect
Fiducial seqs.	$\{(\cdot),\ (H),\ (H,S,H),\ (S,H,S)\}$	
Training exps.	100 random RB sequences of increasing length n	$n = 40,\dots 60$
Testing exps.	100 random RB sequences at each of 87 logarithmically spaced n	$n = 10,\dots,252$

the Lindbladian $L = (\mathbb{1}\otimes H - H^{\mathrm{T}}\otimes\mathbb{1})$. We perturb the evolution time slightly to write

$$\begin{aligned} G_H(\delta\theta) &= e^{i\left(\frac{\pi}{2}+\delta\theta\right)(H\otimes\mathbb{1}-\mathbb{1}\otimes H)} \\ &= \cos^2(\delta\theta)H\otimes H + \sin^2(\delta\theta)\mathbb{1}\otimes\mathbb{1} + \\ &\qquad \frac{i}{2}\sin(2\delta\theta)\left(\mathbb{1}\otimes H - H\otimes\mathbb{1}\right). \end{aligned} \tag{30}$$

An S gate is simply a $R_z\left(\frac{\pi}{2}\right)$ gate, so in line with the previous examples, we choose a distribution $R_z\left(\frac{\pi}{2}+\delta\theta\right)$ where $\delta\theta$ is normally distributed with mean 0 and variance σ_θ^2.

We then add Ginibre noise to both H and S by sampling random super-operators from the BCSZ distribution, such that the sampled super-operators will have the form

$$G_H \to (1-\epsilon)G_H\left(\delta\theta_H\right) + \epsilon\Lambda_H \quad \text{and} \tag{31}$$

$$G_S \to (1-\epsilon)R_z\left(\frac{\pi}{2}+\delta\theta_S\right) + \epsilon\Lambda_S. \tag{32}$$

For the presented example, we chose $\epsilon = 10^{-3}$, and $\delta\theta_H, \delta\theta_S \in \mathcal{N}(0, \sigma^2 = 0.0015)$. With respect to the choice of σ^2, it can be shown that a channel overrotated by $\delta\theta$ has fidelity $F = \frac{2}{3}+\frac{1}{3}\cos(2\delta\theta)$. Fidelities on the order of 0.999 are typical of qubits today, and so assuming that $\delta\theta$ corresponds roughly to the standard deviation, we choose $\sigma^2 = 0.0015$. As we are assuming the addition of Ginibre noise, the actual fidelity of our operations will be slightly lower than this.

To generate data, we chose a 'true' version of G_H and G_S by sampling from the prior. The sampled parameters, as listed in the supplementary materials, are

$$\delta\theta_H = -0.007798, \quad \delta\theta_S = -0.047391, \tag{33}$$

$$\Lambda_H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.435103 & -0.120449 & -0.297836 & 0.062722 \\ -0.314789 & 0.032982 & -0.089239 & 0.080124 \\ 0.188424 & 0.101214 & -0.284711 & 0.142465 \end{pmatrix}$$

$$\Lambda_S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -0.256911 & 0.53382 & 0.265858 & 0.104777 \\ -0.018402 & -0.178172 & 0.565879 & 0.061297 \\ -0.187707 & -0.349921 & -0.279835 & 0.450564 \end{pmatrix}$$

where the super-operators Λ_H, Λ_S are expressed in the Pauli basis.

With this prior distribution, we initialized a cloud of 10000 particles. Bayesian inference was performed to learn G_H and G_S by training with 100 RB sequences of length 40 to 60, using an equal number of sequences at each length. We then tested the model using 87 sequence lengths logarithmically spaced from the range from 10 to 252, using 100 random sequences at each length. For each particle in the posterior distribution, we compute the survival probability for each sequence (the same set of sequences was used for each particle). For each particle we can then fit to a curve of the form $P(m) = (A-B)p^m + B$ to obtain the traditional set of RB fit parameters. The parameters A, B are constrained to be between 0 and 1, and p to be between -0.5 and 1. The fit is a least-squares fit weighted by variance, since at every sequence length the survival probability is averaged over 100 different sequences.

The mean survival probability is shown in Figure 8. At each length, the mean is computed first for each particle over the set of 100 sequences, and then a weighted average over these means is taken using the particle weights in the posterior distribution. Since each particle yields a set of (A, B, p), we can also compute the weighted mean of these parameters, shown as the solid blue line in Figure 8. The mean fit parameters are $(A, B, p) =$

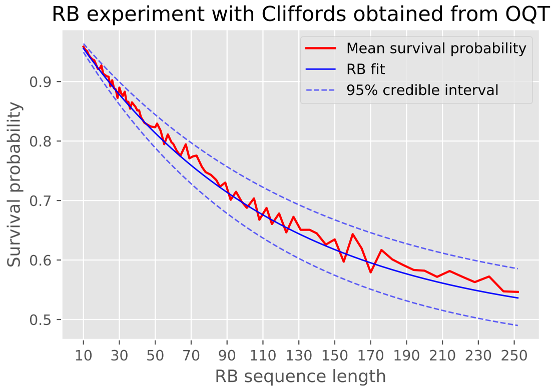

Figure 8: RB decay curve for our learned Clifford group, with 95% credible interval. The survival probability of the y-axis represents the average over RB sequences of a fixed length; the mean survival probability on the plot is the survival probability at length m averaged over 100 experiments, on which we then take the weighted average over the full posterior distribution. We fit a curve for each particle, and display here the curve that is the weighted average of the fit parameters. The fit has the form $P(m) = (A - B)p^m + B$, with mean parameters $(A, B, p) = (0.999916, 0.481494, 0.991119)$, and thus average gate set fidelity 0.995560. The 'true' average gate set fidelity falls within the computed credible interval of $[0.995304, 0.996115]$.

$(0.999916, 0.481494, 0.991119)$ The mean value of p can be used to compute an average gate set fidelity of $(1+p)/2 = 0.995560$. Using the same testing experiments, we can compute the RB decay rate for the 'true' gate set that generated the data. We obtain a 'true' value of 0.995337.

We can plot the 95% credible interval over all RB parameters using the Bonferroni correction [61, 62]. This interval is shown in Figure 8 as dotted lines. We obtain for p the interval $[0.990608, 0.992230]$, corresponding to fidelities of $[0.995304, 0.996115]$, which neatly contains the 'true' value of 0.995337. Details and additional plots are available in the supplementary materials.

6 Quantum mechanics in an operational representation

The previous examples showed a wide range of different characterization tasks that can be implemented within an operational representation. However, a further foundational question is what the limits of this formalism are, and whether or not all of quantum mechanics can be understood within our formalism. One of the fundamental challenges about the gateset tomography model that we inherit is that the description of the gateset is entirely discrete. That is to say that the device in question contains a set of buttons. However, quantum dynamics is naturally continuous. Therefore, understanding quantum dynamics within the language of GST is challenging. Here, we will show that we can think of such continuous dynamics as yielding an equation of motion for the gateset of a quantum system. This not only shows that operational GST is general enough to describe all of quantum mechanics, but also provides a new way of modelling quantum dynamics using only gauge-independent parameters (i.e. observable quantities) and thereby eschewing the use of unobservable quantities such as quantum state operators which appear in conventional treatments of quantum dynamics.

In particular, to allow arbitrary quantum dynamics it is convenient to think now of our operational representation as being an explicit function of time. We assume here for simplicity that the gates, fiducials and measurements all are given by time-independent sequences. As an example, let us consider the case where $\partial_t|\rho(t)\rangle\!\rangle = \mathcal{L}|\rho(t)\rangle\!\rangle$, where $\mathcal{L}$ is the Lindbladian super-operator and $|\rho(t)\rangle\!\rangle$ is the initial state evaluated for the system at time t.

The equation of motion for the operational representation of the gate set is then

$$\begin{aligned}\partial_t\tilde{E}_i(t) &= \langle\!\langle E|F_i\partial_t|\rho(t)\rangle\!\rangle = \langle\!\langle E|F_i\mathcal{L}|\rho(t)\rangle\!\rangle,\\ \partial_t\tilde{F}_{ij}(t) &= \langle\!\langle E|F_iF_j\partial_t|\rho(t)\rangle\!\rangle = \langle\!\langle E|F_iF_j\mathcal{L}|\rho(t)\rangle\!\rangle, \quad (34)\\ \partial_t\tilde{G}^{(k)}_{ij}(t) &= \langle\!\langle E|F_iG_kF_j\partial_t|\rho(t)\rangle\!\rangle = \langle\!\langle E|F_iG_kF_j\mathcal{L}|\rho(t)\rangle\!\rangle,\end{aligned}$$

A challenge with this representation is that its evaluation relies on objects that we do not know *a priori* and are not related (directly) to observed quantities since the expectation values of $\mathcal{L}$ are not assumed to be known in the operational representation. In part, this has to do with the way that we have chosen to represent $\mathcal{L}$. In the following, let us assume that there exist coefficients α_ℓ such that

$$\mathcal{L} = \sum_\ell \alpha_\ell F_\ell. \qquad (35)$$

These values of α_ℓ can further be learned empirically using the operational representation. Let us assume that we empirically measure by choosing $\delta \ll 1$ and taking $\partial_t\tilde{E}(t) \approx (\tilde{E}(t+\delta) - \tilde{E}(t))/\delta$. If we then take the resultant vector to be $\dot{\tilde{\mathbf{E}}}(t)$, $\tilde{\mathbf{F}}(t)$ to be the matrix representation of the $\tilde{F}_{ij}(t)$ tensor and take $\boldsymbol{\alpha}$ to be the unknown matrix of coefficients for the Lindbladian using (35) then if $\tilde{\mathbf{F}}$ is an invertible matrix then

$$\dot{\tilde{\mathbf{E}}}(t) = \tilde{\mathbf{F}}(t)\boldsymbol{\alpha} \Rightarrow \boldsymbol{\alpha} = \tilde{\mathbf{F}}^{-1}(t)\dot{\tilde{\mathbf{E}}}(t). \qquad (36)$$

Thus such a representation can be learned if $\tilde{\mathbf{F}}(t)$ is an invertible matrix. If not, a least-squares approximation can be found by applying the Moore-Penrose pseudoinverse in its place. Of course, this merely proves the existence of a solution (or a least-squares solution) for the coefficients of the Lindbladian as a function of the Fiducials. In practice, Bayesian methods such as the ones considered here and elsewhere may be of great use for both learning and quantifying the uncertainty in the model Lindbladian.

Given a set of coefficients for the Lindbladian the first order system of equations that governs the evolution can be expressed as

$$\begin{aligned} \partial_t \tilde{E}_i(t) &= \sum_\ell \alpha_\ell \langle E|F_i F_\ell|\rho(t)\rangle, \\ \partial_t \tilde{F}_{ij}(t) &= \sum_\ell \alpha_\ell \langle E|F_i F_j F_\ell|\rho(t)\rangle, \qquad (37) \\ \partial_t \tilde{G}^{(k)}_{ij}(t) &= \sum_\ell \alpha_\ell \langle E|F_i G_k F_j F_\ell|\rho(t)\rangle, \qquad (38) \end{aligned}$$

As we can see the derivatives of $\widetilde{E}_i(t)$ depend on the values of $\widetilde{F}_{ij}(t)$ but the derivatives of $\widetilde{F}_{ij}$ and $\widetilde{G}^{(k)}_{ij}$ depend on expectation values of cubic functions of the fiducials. Thus we can solve these equations, but doing so may require more information in some cases. Below we consider two important cases. The first is where the set of fiducial super-operators is not closed under multiplication and the second is where the group is closed and and consists of at most quadratic polynomials in the fiducials.

6.1 Dynamics for infinite sets of fiducials

As a first example of how the dynamics of the operational representation works, consider the case where the fiducial super-operators form an infinite group wherein the group product is given by super-operator multiplication. In this case, we cannot assume any structure to the fiducials that will cause products of them to contract to a finite set of super-operators.

If we have such a model then the dynamics can again be written in terms of a set of observables, however the set that needs to be measured becomes larger in this setting. In particular, we extend the definition of the $\widetilde{E}$ and $\widetilde{G}$ tensors such that

$$\begin{aligned} \tilde{E}_{i_1,\ldots,i_p}(t) &= \langle E|F_{i_1}\cdots F_{i_p}|\rho(t)\rangle. \\ \tilde{G}^{(k)}_{ij_1,\ldots,j_p}(t) &= \langle E|F_i G_k F_{j_1}\cdots F_{j_p}|\rho(t)\rangle. \qquad (39) \end{aligned}$$

Under these assumptions the dynamics of the operational representation of the gate set takes the form of a driven first order dynamical system.

$$\begin{aligned} \partial_t \tilde{E}_i(t) &= \sum_\ell \alpha_\ell \tilde{F}_{i\ell}(t), \\ \partial_t \tilde{F}_{ij}(t) &= \sum_\ell \alpha_\ell \tilde{E}_{ij\ell}(t), \\ &\vdots \\ \partial_t \tilde{F}_{i_1\ldots i_n}(t) &= \sum_\ell \alpha_\ell \tilde{E}_{i_i\ldots i_n\ell}(t), \\ &\vdots \\ \partial_t \tilde{G}^{(k)}_{ij}(t) &= \sum_\ell \alpha_\ell \tilde{G}^{(k)}_{ij\ell}(t) \\ &\vdots \\ \partial_t \tilde{G}^{(k)}_{ij_1\ldots j_n}(t) &= \sum_\ell \alpha_\ell \tilde{G}^{(k)}_{ij_1\ldots j_n\ell}(t) \\ &\vdots \qquad (40) \end{aligned}$$

Thus the entire dynamics of the gate set can be predicted if the $\tilde{E}$ and $\tilde{G}$ tensors are known in their entirety. This is operationally equivalent to the Schrödinger equation, while eschewing the need for unobservable quantities such as the quantum state. While solving the resultant dynamical equations formally requires knowing an infinite hierarchy of terms to predict future dynamics perfectly, in many cases the super-operators for the fiducials will form a finite group making knowledge of the complete hierarchy of tensors unnecessary.

Finally, in practice the entire hierarchy is not needed in order to accurately estimate the dynamics for all subsequent times from data at a single time given the decomposition of the Lindbladian into a sum of fiducials. We have from Taylor's theorem and Stirling's approximation that

$$\begin{aligned} \left|\tilde{E}_i(t+\Delta) - \sum_{j=0}^{K} \frac{\partial_t^j \tilde{E}_i(t)\Delta^j}{j!}\right| &\le \frac{(\sum_\ell |\alpha_\ell|\Delta)^{K+1}}{(K+1)!} \\ &\le \left(\frac{(\sum_\ell |\alpha_\ell|\Delta)}{K+1}\right)^{K+1}. \qquad (41) \end{aligned}$$

Thus by solving this equation for the value of K that yields error ϵ we find that a sufficient value of K is

$$K = \left\lfloor \frac{\ln(1/\epsilon)}{\text{LambertW}\left(\frac{\ln(1/\epsilon)}{(\sum_\ell |\alpha_\ell|\Delta)}\right)} \right\rfloor \in O\left(\frac{\ln(1/\epsilon)}{\ln(\ln(1/\epsilon))}\right), \qquad (42)$$

if $\Delta \le \sum_\ell |\alpha_\ell|$. Thus the total number of terms needed to simulate the dynamics for a short time step with error at most ϵ. varies logarithmically with the error tolerance. Each such term can be approximated using Monte-Carlo sampling such that the estimate of the derivatives is at most ϵ using a number of samples that scales as $O(\mathrm{poly}(1/\epsilon))$ and therefore even in the case where the algebra does not close the dynamics can be simulated using a small number of observables. It should be noted that in the event that the fiducials form a closed group that this scaling improves exponentially Monte-Carlo sampling is no longer required to estimate the derivatives. This shows that under reasonable assumptions the operational representation can also be used to describe the dynamics of a quantum system that we can probe experimentally using a set of fiducial operations and gates. Hence, while inspired by problems of characterization in quantum systems, much broader classes of quantum dynamical problems can also be discussed using our formalism while only making reference to observable quantities.

6.2 Dynamics for closed sets of fiducials

Next let us consider a simpler case where the set of fiducial super-operators is closed under multiplication. Specifically, let $S = \{F_i \bigcup F_i F_j\}$ be the set of all monomials and binomials in the fiducials. Next because the set is closed under multiplication there exists a function g such that for any s_i and s_j in S there exists $s_{g(i,j)}$ such that $s_i s_j = s_{gf(i,j)}$. Also for simplicity, assume that the sets are laid out in lexicographic ordering such that $s_1 = F_1, s_2 = F_2, \ldots$. It then follows that if we use the fact that the set is closed then the equations of motion for the operational representation greatly simplify to the following finite system of equations

$$\begin{aligned}
\partial_t \tilde{E}_i(t) &= \sum_\ell \alpha_\ell \langle\!\langle E|F_i F_\ell|\rho(t)\rangle\!\rangle = \sum_\ell \alpha_\ell \widetilde{F}_{i\ell}(t),\\
\partial_t \tilde{F}_{ij}(t) &= \sum_\ell \alpha_\ell \langle\!\langle E|s_{g(i,g(j,\ell))}|\rho(t)\rangle\!\rangle,\\
\partial_t \tilde{G}^{(k)}_{ij}(t) &= \sum_\ell \alpha_\ell \langle\!\langle E|F_i G_k F_j F_\ell|\rho(t)\rangle\!\rangle = \sum_\ell \widetilde{G}^{(k)}_{ij\ell}(t),\\
\partial_t \tilde{G}^{(k)}_{ij_1 j_2}(t) &= \sum_\ell \alpha_\ell \langle\!\langle E|F_i G_k s_{g(j_1,g(j_2,\ell))}|\rho(t)\rangle\!\rangle \qquad (43)
\end{aligned}$$

These equations can, in many cases be solved directly without having to truncate (as was done in the infinite case considered above). Also, because of the lack of a curse of dimensionality the resulting equations can be solved within error ϵ using $O(\mathrm{polylog}(1/\epsilon))$ operations via existing differential equations solvers. For this reason, cases where the fiducial operators form a closed (or approximately closed) set under multiplication can greatly simplify the equations of motion. However, it should be noted that such cases are highly restrictive and, for example, preclude the inclusion of depolarizing noise or similar effects because such noise models will typically lead to a fiducial set that is not closed under multiplication. For such situations, truncating the infinite dynamics at finite order may be preferable.

6.3 Generator Learning

A further observation is that because the dynamics considered above can be represented as a set of coupled first-order differential equations, we can use this representation to think of generalize the notion of Hamiltonian learning beyond the framework originally proposed. In particular, consider the dynamics in (43). Let us consider a concatenation of all such terms in (43) of the form

$$\Psi(t) = [\tilde{\boldsymbol{E}}(t), \tilde{\boldsymbol{F}}(t), \tilde{\boldsymbol{G}}^{(k)}_{ij}(t), \tilde{\boldsymbol{G}}^{(k)}_{ij\ell}(t)]^T, \qquad (44)$$

where we explicitly include the indices of $\boldsymbol{G}$ above to differentiate the rank 2 and rank 3 $\boldsymbol{G}$ tensors. We then have from the theory of differential equations and the fact that (43) is a homogeneous first-order differential equation that there exists a generator $\boldsymbol{K}$ such that for any condition $\Psi(0)$,

$$\partial_t \Psi(t) = K\Psi(t), \qquad \Psi(t) = e^{Kt}\Psi(0). \qquad (45)$$

This means that we can also infer dynamical models for a gateset using Bayesian inference. In particular, if we have an initial description of our gateset $\Psi(0)$ then evolve some time t and after applying a gate sequence then we would have that the posterior distribution can be expressed as

$$P(K|E) = \frac{P(E|K;\Psi(0)G_0,\ldots,G_{N-1},t)P(K)}{\int P(E|K;\Psi(0)G_0,\ldots,G_{N-1},t)P(K)\mathrm{d}K} \qquad (46)$$

This model ends up assuming that the time required for the gate sequence to be implemented is negligible compared to the dynamical timescale for the gateset. In the event that the timescales are comparable, then K only becomes the instantaneous generator of time-displacements for the gateset and the result will become an ordered operator exponential rather than the simple operator exponential given above.

The key point behind these observations is that techniques that are more reminiscent of quantum Hamiltonian learning (such as in [44]) can also be included within our operational representation. This not only shows that the framework is broader than it may have first appeared but also that we can apply the same ideas employed in that literature in order to infer models for the dynamics of a gateset. This allows some forms of non-Markovian noise to be incorporated in our models without leaving the operational representation.

7 Conclusions

We have demonstrated a framework for quantum tomography in which we can represent many other characterization tasks. Working with a gauge-independent representation of the system, we can learn its behavior from experimental data and predict the outcomes of future experiments. OQT gives us the freedom to incorporate prior information computationally (without any physical experiments). Future improvements to OQT involve the extension to two-qubit operations, as well as allowing for buttons to be held down for arbitrary duration (i.e. time-dependent operations).

Acknowledgments

CG thanks Chris Ferrie for teaching him the contents of Appendix A over the years. ODM is grateful to the quantum group at Microsoft for numerous visits and the opportunity to participate in an internship. TRIUMF receives funding via a contribution through the National Research Council of Canada. IQC is supported in part by the Government of Canada and the Province of Ontario.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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 DE-NA0003525. This research was funded, in part, by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research Quantum Testbed Program. 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.

This research was supported by PNNL's Quantum Algorithms, Software, and Architectures (QUASAR) LDRD Initiative.

References

[1] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, "Elucidating reaction mechanisms on quantum computers," Proceedings of the National Academy of Sciences , 201619152 (2017).

[2] C. Ferrie, "Self-Guided Quantum Tomography," Physical Review Letters **113**, 190404 (2014).

[3] J. B. Altepeter, D. Branning, E. Jeffrey, T. C. Wei, P. G. Kwiat, R. T. Thew, J. L. O'Brien, M. A. Nielsen, and A. G. White, "Ancilla-assisted quantum process tomography," Phys. Rev. Lett. **90**, 193601 (2003).

[4] C. Granade, J. Combes, and D. G. Cory, "Practical Bayesian tomography," New Journal of Physics **18**, 033024 (2016).

[5] R. Blume-Kohout, "Optimal, reliable estimation of quantum states," New J. Phys. **12**, 043034 (2010).

[6] F. Huszár and N. M. T. Houlsby, "Adaptive Bayesian quantum tomography," Physical Review A **85**, 052120 (2012).

[7] D. C. McKay, A. W. Cross, C. J. Wood, and J. M. Gambetta, "Correlated randomized benchmarking," (2020), arXiv:2003.02354 [quant-ph] .

[8] M. Quadeer, M. Tomamichel, and C. Ferrie, "Minimax quantum state estimation under bregman divergence," Quantum **3**, 126 (2019).

[9] P. Cerfontaine, R. Otten, and H. Bluhm, "Self-consistent calibration of quantum gate sets," (2019), arXiv:1906.00950 [quant-ph] .

[10] T. Guff, Y. R. Sanders, N. A. McMahon, and A. Gilchrist, "Decision-making in quantum state discrimination," (2019), arXiv:1906.09737 [quant-ph] .

[11] L. J. Fiderer, J. Schuff, and D. Braun, "Neural-network heuristics for adaptive bayesian quantum estimation," (2020), arXiv:2003.02183 [quant-ph] .

[12] J. M. Lukens, K. J. Law, A. Jasra, and P. Lougovski, "A practical and efficient approach for bayesian quantum state estimation," (2020), arXiv:2002.10354 [quant-ph] .

[13] J. B. Altepeter, D. Branning, E. Jeffrey, T. C. Wei, P. G. Kwiat, R. T. Thew, J. L. O'Brien, M. A. Nielsen, and A. G. White, "Ancilla-assisted quantum process tomography," Phys. Rev. Lett. **90**, 193601 (2003).

[14] 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," Physical Review A **87**, 062119 (2013).

[15] 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," (2013), arXiv:1310.4492 [quant-ph] .

[16] J. P. Dehollain, J. T. Muhonen, R. Blume-Kohout, K. M. Rudinger, J. K. Gamble, E. Nielsen, A. Laucht, S. Simmons, R. Kalra, A. S. Dzurak, *et al.*, "Optimization of a solid-state electron spin qubit using gate set tomography," New Journal of Physics **18**, 103018 (2016).

[17] 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," Nature Communications **8**, 14485 (2017).

[18] K. Rudinger, S. Kimmel, D. Lobser, and P. Maunz, "Experimental demonstration of a cheap and accurate phase estimation," Phys. Rev. Lett. **118**, 190502 (2017).

[19] K. Rudinger, T. Proctor, D. Langharst, M. Sarovar, K. Young, and R. Blume-Kohout, "Probing context-dependent errors in quantum processors," Physical Review X **9**, 021045 (2019).

[20] M. Rol, C. Bultink, T. O'Brien, S. De Jong, L. Theis, X. Fu, F. Luthi, R. Vermeulen, J. de Sterke, A. Bruno, *et al.*, "Restless tuneup of high-fidelity qubit gates," Physical Review Applied **7**, 041001 (2017).

[21] Y. Chen, M. Farahzad, S. Yoo, and T.-C. Wei, "Detector tomography on IBM quantum computers and mitigation of an imperfect measurement," Physical Review A **100**, 052315 (2019).

[22] M. R. Geller, "Rigorous measurement error correction," (2020), arXiv:2002.01471 [quant-ph] .

[23] L. Govia, G. Ribeill, D. Ristè, M. Ware, and H. Krovi, "Bootstrapping quantum process tomography via a perturbative ansatz," Nature communications **11**, 1 (2020).

[24] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, "Demonstration of a parametrically activated entangling gate protected from flux noise," Physical Review A **101**, 012302 (2020).

[25] A. Hughes, V. Schäfer, K. Thirumalai, D. Nadlinger, S. Woodrow, D. Lucas, and C. Ballance, "Benchmarking of a high-fidelity mixed-species entangling gate," (2020), arXiv:2004.08162 [quant-ph] .

[26] M. K. Joshi, A. Elben, B. Vermersch, T. Brydges, C. Maier, P. Zoller, R. Blatt, and C. F. Roos, "Quantum information scrambling in a trapped-ion quantum simulator with tunable range interactions," (2020), arXiv:2001.02176 [quant-ph] .

[27] S. Mavadia, C. Edmunds, C. Hempel, H. Ball, F. Roy, T. Stace, and M. Biercuk, "Experimental quantum verification in the presence of temporally correlated noise," NPJ Quantum Information **4**, 1 (2018).

[28] M. Ware, G. Ribeill, D. Riste, C. Ryan, B. Johnson, and M. P. da Silva, "Experimental pauli-frame randomization on a superconducting qubit," Bulletin of the American Physical Society **62** (2017).

[29] E. Nielsen, R. J. Blume-Kohout, K. M. Rudinger, T. J. Proctor, L. Saldyt, *et al.*, *Python GST Implementation (PyGSTi) v. 0.9*, Tech. Rep. (Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2019).

[30] E. Nielsen, K. Rudinger, T. Proctor, A. Russo, K. Young, and R. Blume-Kohout, "Probing quantum processor performance with pyGSTi," (2020), arXiv:2002.12476 [quant-ph] .

[31] Ł. Rudnicki, Z. Puchała, and K. Zyczkowski, "Gauge invariant information concerning quantum channels," Quantum **2**, 60 (2018).

[32] J. Lin, J. J. Wallman, and R. Laflamme, "Independent state and measurement characterization in quantum computers," (2019), arXiv:1910.07511 [quant-ph] .

[33] A. Doucet and A. M. Johansen, *A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later* (2011).

[34] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97 (1970).

[35] M. Betancourt, "A Conceptual Introduction to Hamiltonian Monte Carlo," (2017), arXiv:1701.02434 [stat] .

[36] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," (2011), arXiv:1111.4246 [cs, stat] .

[37] I. Hincks, J. J. Wallman, C. Ferrie, C. Granade, and D. G. Cory, "Bayesian Inference for Randomized Benchmarking Protocols," (2018), arXiv:1802.00401 [quant-ph] .

[38] A. L. Barker, D. E. Brown, and W. N. Martin, "Bayesian estimation and the Kalman filter," Computers & Mathematics with Applications **30**, 55 (1995).

[39] C. Granade, C. Ferrie, I. Hincks, S. Casagrande, T. Alexander, J. Gross, M. Kononenko, and Y. Sanders, "QInfer: Statistical inference software for quantum applications," Quantum **1**, 5 (2017).

[40] J. Liu and M. West, "Combined parameter and state estimation in simulation-based filtering," in *Sequential Monte Carlo Methods in Practice*, edited by D. Freitas and N. Gordon (Springer-Verlag, New York, 2001).

[41] M. A. Nielsen and I. Chuang, "Quantum computation and quantum information," (2002).

[42] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," IEEE Transactions on Information Theory **51**, 2664 (2005).

[43] R. Kueng and C. Ferrie, "Near-optimal quantum tomography: estimators and bounds," New Journal of Physics **17**, 123013 (2015).

[44] C. E. Granade, C. Ferrie, N. Wiebe, and D. G. Cory, "Robust online Hamiltonian learning," New Journal of Physics **14**, 103013 (2012).

[45] W. Bruzda, V. Cappellini, H.-J. Sommers, and K. Życzkowski, "Random quantum operations," Physics Letters A **373**, 320 (2009).

[46] E. Magesan, J. M. Gambetta, and J. Emerson, "Characterizing quantum gates via randomized benchmarking," Physical Review A **85** (2012).

[47] J. Emerson, R. Alicki, and K. Życzkowski, "Scalable noise estimation with random unitary operators," J. Opt. B Quantum Semiclass. Opt. **7**, S347 (2005).

[48] 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).

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

[50] E. Magesan, J. M. Gambetta, and J. Emerson, "Scalable and robust randomized benchmarking of quantum processes," Phys. Rev. Lett. **106**, 180504 (2011).

[51] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, "Characterizing universal gate sets via dihedral benchmarking," Phys. Rev. A **92**, 060302 (2015).

[52] A. W. Cross, E. Magesan, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, "Scalable randomised benchmarking of non-Clifford gates," NPJ Quantum Inf. **2**, 16012 (2016).

[53] W. G. Brown and B. Eastin, "Randomized benchmarking with restricted gate sets," Phys. Rev. A **97**, 062323 (2018).

[54] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, "Real randomized benchmarking," Quantum **2**, 85 (2018).

[55] J. Helsen, J. J. Wallman, S. T. Flammia, and S. Wehner, "Multiqubit randomized benchmarking using few samples," Phys. Rev. A **100**, 032304 (2019).

[56] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, "Three qubit randomized benchmarking," Phys. Rev. Lett. **122**, 200502 (2019).

[57] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, "Investigating the limits of randomized benchmarking protocols," Phys. Rev. A **89**, 062321 (2014).

[58] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, "What randomized benchmarking actually measures," Phys. Rev. Lett. **119**, 130502 (2017).

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

[60] T. J. Proctor, A. Carignan-Dugas, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, "Direct randomized benchmarking for multiqubit devices," Physical Review Letters **123**, 030503 (2019).

[61] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," Istituto Superiore di Scienze Economiche e Commerciali di Firenze **8**, 3 (1936).

[62] O. J. Dunn, "Multiple comparisons among means," Journal of the American Statistical Association **56**, 52 (1961).

[63] J. Kiefer, "On wald's complete class theorems," Ann. Math. Statist. **24**, 70 (1953).

[64] V. Al Osipov, H.-J. Sommers, and K. Życzkowski, "Random bures mixed states and the distribution of their purity," Journal of Physics A: Mathematical and Theoretical **43**, 055302 (2010).

[65] B. H. Fong and S. T. Merkel, "Randomized Benchmarking, Correlated Noise, and Ising Models," (2017), arXiv:1703.09747 [quant-ph] .

[66] C. Granade, C. Ferrie, and D. G. Cory, "Accelerated randomized benchmarking," New Journal of Physics **17**, 013042 (2015).

[67] R. W. Heeres, P. Reinhold, N. Ofek, L. Frunzio, L. Jiang, M. H. Devoret, and R. J. Schoelkopf, "Implementing a universal gate set on a logical qubit encoded in an oscillator," Nature Communications **8**, 94 (2017).

[68] D. Steel, "Bayesian Confirmation Theory and The Likelihood Principle," Synthese **156**, 53 (2007).

[69] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, "Superconducting quantum circuits at the surface code threshold for fault tolerance," Nature **508**, 500 (2014).

A Review of Bayesian estimators

In this Appendix, we provide a brief review of estimation theory as applied to Bayesian inference. In doing so, it is convenient to first consider estimation more generally. Suppose that there is some vector $\boldsymbol{x} \in \mathcal{X}$ of parameters that we would like to learn given some data $D \in \mathcal{D}$, where $\mathcal{X}$ is the set of feasible values for $\boldsymbol{x}$, and where $\mathcal{D}$ is the set of data we could have possibly obtained. Then, we will say that any function $\hat{\boldsymbol{x}}(\cdot) : \mathcal{D} \to \mathcal{X}$ which accepts data and returns estimates is an *estimator*.

For example, given any $\boldsymbol{x}_0 \in \mathcal{X}$, the constant function $\hat{\boldsymbol{x}}(D) = \boldsymbol{x}_0$ is an estimator that disregards all evidence in favor of returning $\boldsymbol{x}_0$. Clearly, while this is a valid estimator, it is not a very *good* one to use in practice. Our task in estimation theory is then to recommend a particular estimator that is desirable according to some set of practical considerations. We may, for example, want an estimator that incurs as little error as possible.

We can formalize this desire by introducing a function $L : (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}^+$ such that $L(\hat{\boldsymbol{x}}, \boldsymbol{x})$ is the *loss* that we incur if we return $\hat{\boldsymbol{x}}$ as our estimate when the true value is $\boldsymbol{x}$. For example, if we are estimating a single real number ($\mathcal{X} = \mathbb{R}$), then we may choose the squared error $L(\hat{x}, x) = (\hat{x} - x)^2$ as our loss. More generally, for $\mathcal{X} = \mathbb{R}^d$ for $d \in \mathbb{N}$, the quadratic loss $L_{\boldsymbol{Q}}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = (\hat{\boldsymbol{x}} - \boldsymbol{x})^{\mathrm{T}} \boldsymbol{Q} (\hat{\boldsymbol{x}} - \boldsymbol{x})$ is a well-defined loss function for any positive definite matrix $\boldsymbol{Q}$.

Once we have decided upon a loss function, we can then reason about what losses we may incur in a given experiment using a particular estimator. To do so, we first need to extend our definition of loss from estimates to estimators by taking the average over all possible data sets that an estimator could take as input. Concretely, given a loss function L, define the *risk* $R : (\mathcal{D} \to \mathcal{X}) \to \mathbb{R}^+$ of an estimator as

$$R(\hat{\boldsymbol{x}}, \boldsymbol{x}) := \mathbb{E}_{D \sim \Pr(D|\boldsymbol{x})}[L(\hat{\boldsymbol{x}}(D), \boldsymbol{x})]. \tag{47}$$

The risk implicitly defines a multi-objective optimization problem, in that an estimator that works well for a particular ground truth need not work well more generally. At an extreme, the constant estimator $\hat{\boldsymbol{x}}(D) = \boldsymbol{x}_0$ works beautifully well when $\boldsymbol{x} = \boldsymbol{x}_0$. We thus at a minimum want an estimator that minimizes the risk that we incur in some case of interest. To formalize this notion, we say that an estimator $\hat{\boldsymbol{x}}(\cdot)$ is *dominated* by an estimator $\hat{\boldsymbol{x}}'(\cdot)$, if for all $\boldsymbol{x}$, $R(\hat{\boldsymbol{x}}, \boldsymbol{x}) \geq R(\hat{\boldsymbol{x}}', \boldsymbol{x})$, and if there exists some $\boldsymbol{x}$ for which this inequality is strict. Put differently, an estimator dominates another estimator if it is less risky in all circumstances, such that there is no decision-theoretic basis for preferring the dominated estimator. An estimator which is not dominated by any other estimator is said to be *admissible*.

From a Bayesian perspective, however, we are generally most interested in minimizing what we expect the risk to be given our experience with a system so far. We can make this precise by taking the expectation value of the risk with respect to a prior distribution to obtain the *Bayes risk* of an estimator,

$$r(\hat{\boldsymbol{x}}) := \mathbb{E}_{\boldsymbol{x} \sim \Pr(\boldsymbol{x})}[R(\hat{\boldsymbol{x}}, \boldsymbol{x})]. \tag{48}$$

The unique estimator minimizing the Bayes risk for a particular loss function is called the *Bayes estimator* for that loss,

$$\hat{\boldsymbol{x}}_{\text{Bayes}} := \arg\min_{\hat{\boldsymbol{x}}(\cdot)} r(\hat{\boldsymbol{x}}(\cdot)). \tag{49}$$

By construction, the Bayes estimator is admissible: any estimator that dominates the Bayes estimator would have a strictly smaller Bayes risk. Under fairly weak conditions [63], however, we can conclude the converse as well, namely that every admissible estimator is the Bayes estimator for a particular prior distribution.

In full generality, computing the Bayes estimator for a particular loss function requires minimizing over functions of all data sets, which is not feasible or practical. Some loss functions, however, allow for much more efficiently computing Bayes estimators. In particular, *Bregman divergences* are loss functions which can be written as the difference between a convex function and its first-order Taylor expansion. If a loss function is Bregman, then the celebrated theorem of [42] shows that

$$\hat{\boldsymbol{x}}_{\text{Bayes}}(D) = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}|D]. \tag{50}$$

That is, the posterior mean of $\boldsymbol{x}$ is the Bayes estimator for any Bregman divergence.

Many practically relevant loss functions are Bregman divergences, including the squared error, quadratic loss, and Kullback–Liebler divergence. Thus, the posterior mean gives us a method of efficiently computing admissible estimators that minimize the average error we incur in inference procedures. As we saw in Section 3.2, the posterior mean can be efficiently computed using particle filtering, giving us a practical method for reporting Bayes estimates.

B Quantum state tomography

In traditional quantum state tomography, we seek to learn an unknown state using a set of measurements that are presumed to be perfectly known. Naively performing this task, however, leads to estimates that are not self-consistent. We provide a demonstration of this by performing OQT on unknown rebits, i.e. qubits with no y-components.

As in previous examples, the first step is to phrase the problem in terms of our operational formalism. We will consider the case where our box again has 4 buttons: two SPAM buttons, and ones that we believe perform $R_x\left(\frac{\pi}{2}\right)$ (denoted b_{R_x}) and $R_y\left(\frac{\pi}{2}\right)$ (denoted b_{R_y}). We add uncertainty to our rotation buttons by setting the priors for $R_x\left(\frac{\pi}{2}\right)$ and $R_y\left(\frac{\pi}{2}\right)$ to be over-rotations with a mean of 0 and a variance of 10^{-3}. We also add depolarization to the rotation gates, with strength $p \in \mathcal{U}(0, 0.1)$.

We sample our states from the Ginibre rebit distribution, the uniform distribution over rebit states. Such states are sampled by computing

$$\rho = \frac{XX^\dagger}{\operatorname{Tr}(XX^\dagger)}, \quad X_{ij} \in \mathcal{N}(0,1), \tag{51}$$

where in our case, X is a 2×3 matrix[9]. The rebit states are subject to a small amount of depolarization with strength $p \in \mathcal{U}(0, 0.1)$. We apply similar depolarization to the measurement $E = |0\rangle\langle 0|$. Full details of our parameter specifications are shown in Table 4.

The set of chosen fiducial sequences is $\boldsymbol{f} = \{(\cdot), (R_x), (R_y), (R_x, R_x)\}$. If our buttons were

[9] In the more general case, the Ginibre distribution of $d \times d$ density matrices is sampled by populating a $d \times d$ matrix X with complex values $a + bi$ where both a and b are randomly sampled from $\mathcal{N}(0, 1)$. For random real density matrices, we must sample matrices of size $d \times (d + 1)$ [64].

Table 4: OQT parameter specification for rebit state tomography. Button labels are abbreviated $b_{R_x} \to R_x$ for notational simplicity. State tomography was performed independently for 1000 states (and associated gate sets) sampled from the distributions below. As such, we do not provide examples of the sampled parameters in this case. For all priors, the value of p is the amount of depolarization.

Button label	**Prior**	**Example values**
ρ	Ginibre rebit distribution, Eq.(51), $p \in \mathcal{U}(0, 0.1)$	1000 randomly selected states
R_x	$R_x(\pi/2 + \epsilon),\ \ \epsilon \in \mathcal{N}(0, 10^{-3}),\ \ p \in \mathcal{U}(0, 0.1)$	
R_y	$R_y(\pi/2 + \epsilon),\ \ \epsilon \in \mathcal{N}(0, 10^{-3}),\ \ p \in \mathcal{U}(0, 0.1)$	
E	$1/\sqrt{2}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix},\ \ p \in \mathcal{U}(0, 0.1)$	
Fiducial seqs.	$\{(\cdot),\ \ (R_x),\ \ (R_y),\ \ (R_x, R_x)\}$	
Training exps.	50 randomly chosen products of n fiducials	$n = 1, \ldots, 10$; n increases linearly
Testing exps.	50 randomly chosen products of n fiducials	$n = 5, \ldots, 15$; n increases linearly

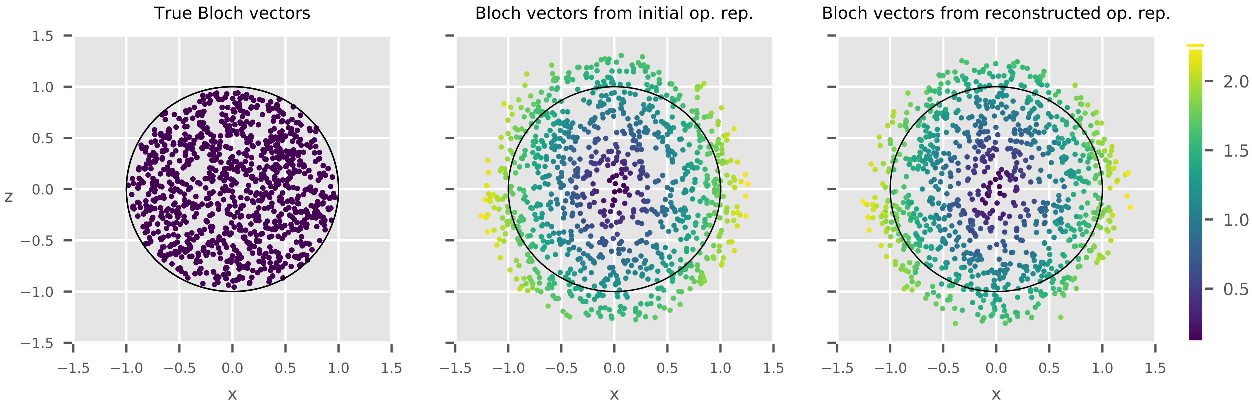

Figure 9: What happens when we perform state tomography 'naively' using the measurement results from noisy gates we had assumed were perfect. (Left) The 1000 initial random states, sampled from the prior. They are rebits, and have only x and z components. (Center) A 'pseudo Bloch circle' constructed by pulling coordinates from the initial operational representation, i.e. fiducial experiment probabilities, as per (52). Points are colored by their distance to the corresponding true states in the left panel. (Right) The same plot as for the middle, but calculating the Bloch coordinates using the posterior mean after performing OQT. See Figure 11 for a histogram of the colored difference before and after reconstruction.

perfect, this set of fiducials provides a set of measurements that is informationally complete in the traditional sense, meaning that the measurements span the entire Bloch sphere. However in practice these will be noisy - our definition of informationally complete thus shifts to whether or not the fiducials produce a well-conditioned $\tilde{F}$; we find that the chosen set is reliable in practice.

In the 'naive' method of performing state tomography, the fiducial sequences and associated probabilities would be directly related to the coordinates on the Bloch sphere (a_x, a_y, a_z):

$$\begin{aligned} a_x &= 2p_x - 1,\ p_x = \tilde{F}_{02} = \mathrm{Tr}\left[|\rho\rangle\langle E|R_y(\pi/2)\right] \\ a_y &= 2p_y - 1,\ p_y = \tilde{F}_{01} = \mathrm{Tr}\left[|\rho\rangle\langle E|R_x(\pi/2)\right] \quad (52) \\ a_z &= 2p_z - 1,\ p_z = \tilde{F}_{00} = \mathrm{Tr}\left[|\rho\rangle\langle E|\right] \end{aligned}$$

In the remainder of this section, we will demonstrate the consequences of this naive method.

We performed state tomography with OQT independently on 1000 random states. In Figure 9, we have plotted the true Bloch coordinates of the initial states in the left panel. In the middle panel, we see the coordinates obtained from their initial operational representations according to (52). States pulled from the Ginibre ensemble should lie firmly within the boundaries of the Bloch sphere, or circle, in the rebit case. However reconstruction according to (52) produces Bloch coordinates that fall well outside the boundaries. Furthermore, they pick up small y-component, as demonstrated in the first two panels of Figure 10.

For our OQT experiments, we push the state preparation button once, then apply a sequence of randomly selected gate buttons from a minimum length of 1 to a maximum length of 10. We then measure, record the outcome, and repeat 50 cases to form a training corpus. The sequence length steadily increased during training, with the same amount of sequences generated at each length.

We note that Figure 9 illustrates the dangers of 'naive' state tomography in the presence of measure-

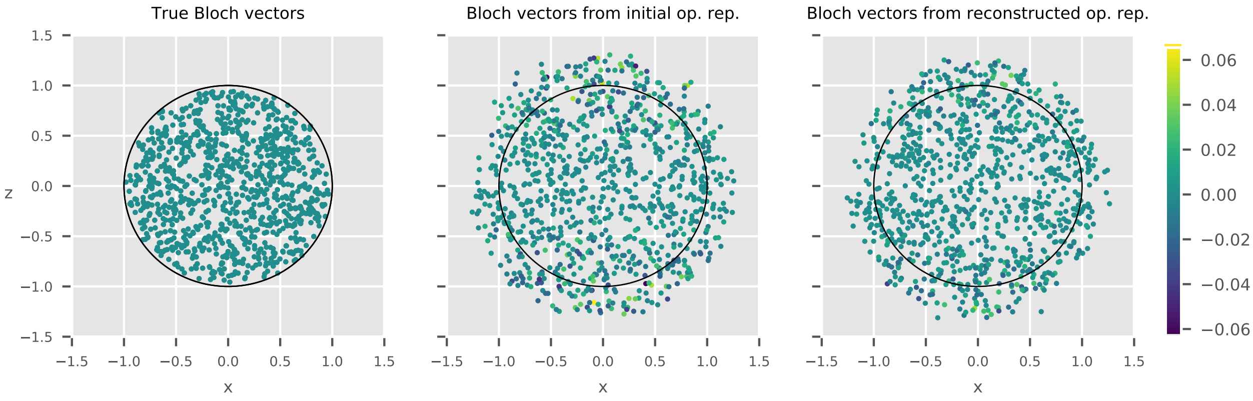

Figure 10: For the same set of states in Figure 9, we color the states according to the y component of the pseudo Bloch vectors. In theory this should always be 0, but we observe here that our naive reconstruction method produces slight deviations both before and after reconstruction. However we note that after reconstruction, the deviation is less, as displayed in the right panel of Figure 11.

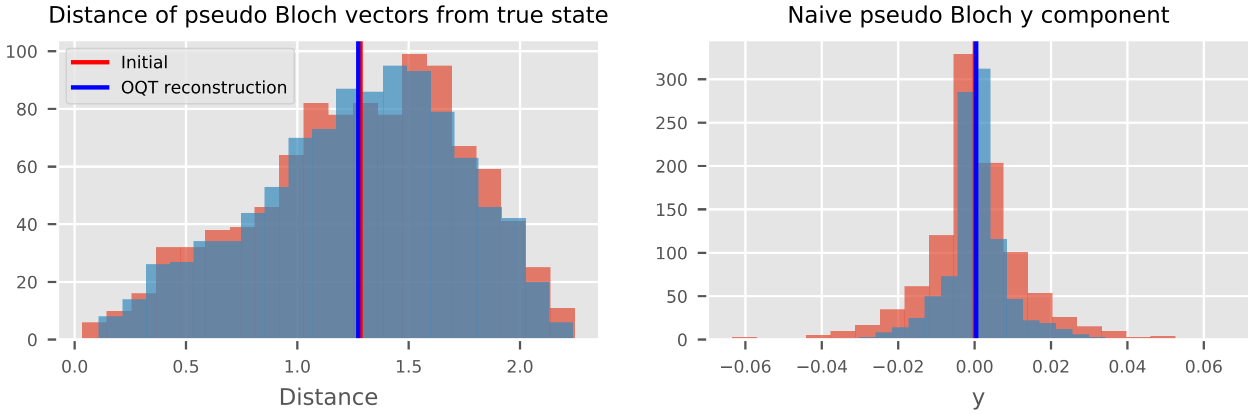

Figure 11: Histograms of pseudo Bloch vector properties before and after performing OQT. Solid lines show the mean of the corresponding distribution. OQT learns these vectors well and produces comparable distributions, but this naive method of tomography nevertheless leads to a noticeable y component added to many of the rebits.

ment errors. For each of the hypotheses shown in the middle and right-hand plots of Figure 9, if a naive tomographer were to take an infinite amount of data from a system described by that hypothesis and then reconstruct the initial state ρ, they would correctly conclude either that their data was impossible, or that ρ lies outside the Bloch sphere entirely. Put differently, if one assumes that the measurement sequences used in a state tomography experiment are ideal, then naive state tomography will return absurd results *even* in the limit of infinite data.

C Details on randomized benchmarking

In this appendix, we discuss the advantages of performing randomized benchmarking within an operational framework.

C.1 Using operational formalisms to perform randomized benchmarking

Magesan *et al.* [46] derived that A and B contain information about the state preparation and measurement errors incurred by a randomized benchmarking experiment. They are expressed analytically as

$$A = \mathrm{Tr}\left(E_\psi \Lambda\left(\rho_\psi\right)\right), \tag{53}$$

$$\text{and } B = \mathrm{Tr}\left(E_\psi \Lambda\left(\frac{\mathbb{1}}{d}\right)\right). \tag{54}$$

A key point here is that traditional RB assumes that Λ is the same for all elements of the Clifford group. However, as we will see, Clifford elements implemented in the GST framework will naturally have different errors, as elements are composed of sequences of H and S of varying lengths.

If the implementations of each Clifford element are perfect, we obtain $A = 1$, $B = 1/2$, and $p = 1$, and so the survival probability is identically 1 for all sequences. However in the worst case, we obtain something essentially depolarized and so $p = 0$, meaning that the curve will immediately decay to $B = 1/2$. Fitting the experimental data to a curve of this form can thus give an idea of the value of p, which in turn can give us an estimate of the average gate fidelity.

Before proceeding, it is helpful to establish that, despite its apparent simplicity, learning figures of merit from randomized benchmarking data is an astonishingly subtle problem that warrants no small amount of caution. Especially given the rigorous demands placed on randomized benchmarking results for application to predicting the success of fault-tolerance, it is of the utmost importance that the results of RB experiments are understood in full recognition of the caveats placed on said results by current experimental and theoretical limitations. For instance, as mentioned above, for instance, the derivation of Magesan *et al.* [46] rests *critically* on the assumption that the noise on each element of a gate set is independent of which element is being considered. While Magesan *et al.* [46] does provide a derivation that attempts to include gate-dependence, later counterexamples have shown that this assumption cannot even be made in a gauge-independent fashion [58] — this implies that the gate-independence assumption cannot be experimentally tested. Later work has shown that the effects of gate-dependence exponentially small effects on randomized benchmarking data [59, 65], but it is still an open question as to how to meaningfully interpret RB data.

Perhaps more pressing still, the original derivation of Magesan *et al.* [46] only derived the *mean* survival probability and not any higher moments. A fitting procedure such as homoscedastic least-squares fitting (the default procedure offered by MATLAB, SciPy, and many other packages, see Appendix C.2 for a review) will thus necessarily give incorrect or misleading answers, as the *variance* over randomized benchmarking data depends both on the variance within each sequence and over shots of that sequence, and on the variance between different sequences. This challenge can be overcome by committing to taking exactly one repetition of each sequence before choosing a new sequence [66], but this is feasible only for a small number of experimental platforms, such as those controlled by custom FPGA firmware [67]. As an alternative solution, one can introduce *nuisance parameters* to track the unknown higher moments and estimate them at the same time as the expectation of interest. A recent proposal of this form was advanced by Hincks *et al.* [37], who introduced a parameterization for RB protocols that includes a distribution at each sequence length that is then sampled using Hamiltonian Monte Carlo, effectively introducing an uncountable number of nuisance parameters in a way that they can be efficiently estimated.

From this perspective, using OQT to analyze randomized benchmarking data provides an explicit and gauge-independent nuisance parameterization that avoids both the interpretational and practical difficulties of drawing inferences from RB data. We can then rely on the procedure of Blume-Kohout *et al.* [17] to synthesize from a final posterior over operational representations RB data of a form that is immediately amenable to analysis by even relatively informal methods such as heteroscedastic least-squares fitting.

C.2 Estimation within randomized benchmarking

In this Appendix, we review the estimation theory underlying randomized benchmarking and summarize some of the most prevalent pitfalls. To do so, we will rely heavily on the Likelihood Principle [68], which informally states that in order to make decisions consistent with experimental observation, we must base our decisions only on the evaluation of a likelihood function at our data, and cannot base our inference on any property of our data that is not expressed in the likelihood. For RB in particular, this consistency requirement forces us to describe our implementations of RB in an operational manner, such that we can write down likelihood functions.

For instance, we recall that as per (26), the Magesan *et al.* [46] model gives us that the mean sequence probability

$$P(m) := (A - B)p^m + B \tag{55}$$

for some parameters $\boldsymbol{y} = (p, A, B)$. This is not yet an operational description, however, as sequence probabilities are *not* observable properties of finite-length experiments [10]. To make an operational description of the Magesan *et al.* [46] model (26), let us be more precise about a description of our experimental procedure. As a prototypical example of such a description, most RB experiments proceed as follows:

1. Perform the following for each $m \in \{m_0, \ldots, m_{M-1}\}$:
    (a) Perform the following N times:
        i. Choose a random sequence $\boldsymbol{s}$.
        ii. Perform the following K times:
            A. Prepare a state ρ.
            B. Apply the sequence $\boldsymbol{s}$
            C. Measure the POVM $\{E, \mathbb{1} - E\}$.

[10] As an amusing aside, this realization implies that the word "observable" in many formulations of quantum mechanics is reserved for those objects which are fail to be observable. It is for this reason that we prefer the more operational description offered by the POVM formulation.

iii. Record the number of times that E was observed in the above loop as $k(\boldsymbol{s})$.

(b) Record the mean of $k(\boldsymbol{s})$ for each $\boldsymbol{s}$ sampled in the above loop as $n(m_i)$.

We recognize the innermost loop as being a sample from the binomial distribution

$$k(\boldsymbol{s}) \sim \mathrm{Bin}(\Pr(E|[\Phi];\boldsymbol{s}), K), \tag{56a}$$

$$\Pr(k|\boldsymbol{s}) = \sum_k \binom{K}{k} p_{\boldsymbol{s}}^k (1-p_{\boldsymbol{s}})^{K-k}, \tag{56b}$$

where we have taken the shorthand $p_{\boldsymbol{s}} := \Pr(E|[\Phi];\boldsymbol{s})$ to denote sequence probabilities of the form considered in the rest of the paper. From the perspective of RB, however, this is problematic, as a sequence probability for the sequence $\boldsymbol{s}$ can in general depend on any element of the operational representation for $[\Phi]$. We may not be able to compute the sequence probability $\Pr(E|[\Phi];\boldsymbol{s})$ given only hypotheses about the RB parameters $\boldsymbol{y}$.

Nonetheless, the Magesan *et al.* [46] model gives us hope that we may still be able to formulate a likelihood function for the entire experiment, even if we cannot do so for each individual sequence within an RB procedure. Following this hope, let us marginalize (56) over the choice of sequence $\boldsymbol{s}$, since we have chosen $\boldsymbol{s}$ randomly at the start of our loop over sequences. Concretely,

$$\begin{aligned}
&\Pr(k|[\Phi];|\boldsymbol{s}|=m_i)\\
&= \mathbb{E}_{\boldsymbol{s}\text{ s.t. }|\boldsymbol{s}|=m_i}\left[\sum_k \binom{K}{k} p_{\boldsymbol{s}}^k (1-p_{\boldsymbol{s}})^{K-k}\right]\\
&= \sum_k \binom{K}{k} \mathbb{E}_{\boldsymbol{s}\text{ s.t. }|\boldsymbol{s}|=m_i}\left[p_{\boldsymbol{s}}^k (1-p_{\boldsymbol{s}})^{K-k}\right].
\end{aligned} \tag{57}$$

Thus, if we wish to compute likelihood functions for K shots at each sequence, we must be able to compute the Kth moment of the distribution of sequence probabilities over all sequences of a given length.

This makes it clear how both the techniques of Granade *et al.* [66] and Hincks *et al.* [37] operate. The former restricts attention to the case in which $K = 1$, such that the needed moment is precisely that given by Magesan *et al.* [46], while the latter introduces additional parameters (formally, nuisance parameters) to track the higher moments of distributions over sequences.

Though both of these approaches are provided along with software implementations, they may be practical constraints that prevent using the $K = 1$ experimental limitation or introducing large numbers of nuisance parameters. In practice, therefore, convenience often demands deviating from statistical principle and exploring what can be done with *ad hoc* methods. For example, least-squares methods are often used in experimental papers to report results from randomized benchmarking observations [69]. In this case, such methods are *ad hoc* in the sense that least-squares fitting requires additional assumptions that are often left implicit.

In particular, if one is attempting to learn the argument x of a function $f(x)$ from samples $y_i = f(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then the least-squares solution can be readily shown to be the maximum likelihood estimator for x. Thus, if a minimum variance unbiased estimator exists for x, it is equal to the least-squares solution. Applying this argument to the RB case thus demands a strong additional assumption be made, namely that

$$n(m_i) \sim \mathcal{N}(P(m), \sigma^2). \tag{58}$$

By using heteroscedastic least-squares fitting, we can relax this assumption such that the variance on each $n(m_i)$ is a function of m_i,

$$n(m_i) \sim \mathcal{N}(P(m), \sigma_i^2). \tag{59}$$

In order to apply heteroscedastic least-squares fitting, we must therefore be able to assume normality, and we must have a way to compute σ_i^2 for each m_i.

In typical experiments, we do not have direct access to such variances. That said, when synthesizing RB data from a posterior over operational representations, something remarkable happens: we can interpret the variance as the mean Bayes risk for the prediction loss over sequences. This interpretation makes it possible to directly compute σ_i^2 from our posterior uncertainty, motivating the use of heteroscedastic least-squares fitting.