GST report for test

Wednesday 10th August, 2016

1 Overview

This report presents a gate-set tomography (GST) analysis of a dataset called "test".

GST characterizes logic operations on a quantum device (e.g., a qubit), by treating it as a black box. This black box is equipped with a small set of "buttons" that apply quantum *gates* to the quantum system inside. One button initializes it, a second button triggers a 2-outcome measurement, and the remaining buttons perform transformations. We avoid assumptions about the device's operation whenever possible. Currently, we assume that:

- the quantum device is a qubit (has a Hilbert space of dimension 2),
- each *gate*, or logic operation, can be represented by a stationary Markov process (a.k.a. "quantum channel").

The core of GST is an algorithm that takes certain inputs, and produces certain outputs. The *input* to GST comprises (1) a list of data, and (2) "target" gate set describing the *ideal* behavior of the device. GST data comprises a list of experiments – each described by the sequence of gates that was applied – and, for each experiment, two integer *counts* stating how often the "plus" and "minus" results were observed. The target gates are used *only* to (a) report how consistent the estimates are with the target, and (b) choose the best *gauge* in which to report the results. GST does not take them into account in its core analysis, and there is no possibility of circularity or other "cheating".

GST's primary output is an estimated *gate set* that models or fits the device's observed behavior. Gate sets are of the form $\{\rho_0, E_0, \{G_k\}\}$, where

- ρ_0 is an estimate of the density matrix in which the device gets initialized,
- $\{E_0, \mathbb{1} - E_0\}$ is an estimate of the POVM describing how it gets measured,
- and each of the G_k is an estimate of the superoperator (quantum process) describing the corresponding gate.

Unless something went wrong (usually it doesn't), the output of GST is the best possible fit to the data. This should also mean that the output is a very accurate description of what happens when you trigger a gate on your device. However, this happy conclusion relies on two assumptions:

1. The experiments were chosen wisely, so that the only gate sets consistent with their results are very close to the true behavior. This is usually true. The main failure mode occurs when you were not able to perform *long* sequences (e.g., because your decoherence rate is very high), in which case accuracy may be limited.
2. The operations you are performing really are stationary (time-independent), Markovian, and acting on a quantum system with the correct Hilbert space dimension. These assumptions define the *model* that GST fits to the data. **They are usually not true!** Quantum operations are usually at least a little bit non-Markovian. In this report (Section 4) we provide extensive self-checks to identify and diagnose violations of the model. If your system *is* visibly non-Markovian, then (a) these checks will probably warn you of it, and (b) the other quantities reported here should be treated with caution – using GST on non-Markovian gates violates the warranty!

This document is organized into three main sections, which address three broad questions.

- Section 2: What inputs did you give GST?
- Section 3: What estimate did GST output, and what does it mean?
- Section 4: How reliable are the results? (How badly was the model violated?)

Section 2 is primarily useful to verify that the inputs were correct. Section 3 is the most important: it presents the raw estimates derived by the GST algorithm, and also provides a variety of derived quantities that may be useful in interpreting what this estimate means.

Section 4 is dedicated to summarizing how well the model imposed by GST was able to fit the data, relative to what is expected of a "good" model. This is *not* related to "How close is the GST estimate to the target gates?", which is addressed in Section 3. It is also not the same as "How large are the error bars on the GST estimate?", which is a good question that is addressed in section 3 when this report is generated with the confidence intervals option turned on. Instead, Section 4 is intended to tell you whether (a) you should take the GST estimate at face value, or (b) it should be treated skeptically because *no* gate set was capable of fitting the data.

Finally, appendices may be present (depending on which options were chosen when this report was generated). Appendices present more detailed debugging information, elaborating on the goodness-of-fit metrics presented in Section 4.

2 Input Summary

The input for this GST analysis comprised: (1) a target gate set (see Tables 1-2); and (2) a dataset called "test".

2.1 Target Gate set

The target gate set describes the ideal initial state (density matrix), measurement (POVM effect), and gate operations (superoperators). Typically, density matrices and POVM effects are represented as square $d \times d$ matrices on a Hilbert space $\mathcal{H}$. In GST, it is often more convenient to represent them as d^2-element vectors in the Hilbert-Schmidt space $\mathcal{B}(\mathcal{H})$ of linear operators on $\mathcal{H}$. Both representations are shown in Table 1. Superoperators are sometimes represented in Choi or Kraus form, but for GST it is more convenient to represent them as square $d^2 \times d^2$ matrices that multiply associatively and act on $\mathcal{B}(\mathcal{H})$. These are shown in Table 2.

These Hilbert-Schmidt space representations require choosing a basis $\{M_i\}$ for $\mathcal{B}(\mathcal{H})$. We use the *Pauli basis*, comprising the four 2×2 Pauli matrices (including the identity $\mathbb{1}$) for $d = 2$. In $d > 2$, we use the analogous Gell-Mann matrices as a basis. The choice of this basis is what is meant when state preparations and measurements are written as vectors and gate operations are written as matrices in the "Pauli basis". Keep in mind that we want to use an orthonormal basis, so the basis matrices are normalized so that $\langle\langle M_i | M_j \rangle\rangle = \mathrm{Tr} M_i^\dagger M_j = \delta_{ij}$. In $d = 2$, this means that the basis matrices are $M_i = \frac{1}{\sqrt{2}}\sigma_i$.

The ideal state prepration and measurement (SPAM) operations for your particular case are given in Table 1. The ideal *logic gate* operations are given, as superoperators written in the Pauli basis, in Table 2.

In most cases, the ideal/target logic gates are reversible unitary rotations. The corresponding superoperators are orthogonal rotations on $\mathcal{B}(\mathcal{H})$. For your convenience, Table 2 also lists (for each logic gate) an axis of rotation [as a vector in $\mathcal{B}(\mathcal{H})$] and an angle of rotation.

2.2 GST Input Data

The most important input to GST is a *dataset* – a list of experimental counts or frequencies, each associated with a *gate sequences*. Gate sequences are also referred to as "gate strings". Each gate sequence defines an experiment, in which you (1) initialize the device, (2) apply the operations specified by the gate sequence, and (3) measure and record the result ("plus" or "minus").

Typically, the gate sequences that appear in the dataset are generated by the following process:

1. A small set of short gate sequences called *germs* are chosen,

Operator	Hilbert-Schmidt vector (Pauli basis)	Matrix
ρ_0	$\begin{matrix}0.7071\\0\\0\\0.7071\end{matrix}$	$\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$
E_0	$\begin{matrix}0.7071\\0\\0\\-0.7071\end{matrix}$	$\begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}$

Table 1: **Target gate set: SPAM (state preparation and measurement) gates**. These are the *ideal* input state (ρ_0) and 'plus' POVM effect E_0 for the device on which we report. SPAM gates are given here both as $d \times d$ matrices, and in "vectorized" form as d^2-dimensional vectors in $\mathcal{B}(\mathcal{H})$. See Table 5 for GST estimates of the actual ρ_0 and E_0 implemented in this experiment.

Gate	Superoperator (Pauli basis)	Rotation axis	Angle
Gi	$\begin{pmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{pmatrix}$	$\begin{matrix}0\\0\\0\\1\end{matrix}$	0π
Gx	$\begin{pmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 0 & -1\\0 & 0 & 1 & 0\end{pmatrix}$	$\begin{matrix}0\\1\\0\\0\end{matrix}$	0.5π
Gy	$\begin{pmatrix}1 & 0 & 0 & 0\\0 & 0 & 0 & 1\\0 & 0 & 1 & 0\\0 & -1 & 0 & 0\end{pmatrix}$	$\begin{matrix}0\\0\\1\\0\end{matrix}$	0.5π

Table 2: **Target gate set: logic gates**. These are the *ideal* (generally unitary) logic gates. Each has a name starting with "G", and is represented as a $d^2 \times d^2$ *superoperator* that acts by matrix multiplication on vectors in $\mathcal{B}(\mathcal{H})$. For each gate, its axis of rotation (in $\mathcal{B}(\mathcal{H})$) and angle of rotation are also given. See Table 7 for GST estimates of the actual logic gates implemented in this experiment.

2. A small set of short *fiducial sequences* are chosen so that, when applied to ρ_0 or E_0, they generate an informationally complete set of states or effects.

3. Each germ is concatenated with itself to form *base sequences* of length approximately $1, 2, 4, 8, \ldots L_{max}$.

4. Each base sequence is sandwiched between every possible pair of fiducial sequences.

The dataset comprises all sandwiched base sequences. A few other short sequences (e.g., those corresponding to the empty base sequence) may also appear.

The fiducial sequences and germs for *this* dataset are given in Table 3. An overview of the information contained in the file you provided for dataset "test" is given in Table 4.

This table also contains one derived quantity, the spectrum of the largest *Gram matrix* that GST could extract from the data. This is included here rather than in the analysis because it is not useful for predictive purposes, and therefore is not part of the estimate. It serves, instead, to tell you something about the quality of the data. More precisely, it tells you about the dimension of the state space that is explored by the fiducial sequences. This should be d^2-dimensional [because the fiducials are intended to explore all of $\mathcal{B}(\mathcal{H})$], and therefore the spectrum listed in Table 4 should (ideally) have exactly d^2 elements that are large and nonzero. In practice, you should see d^2 large elements, and a rapid drop in magnitude thereafter. If fewer than d^2 elements are large, then the fiducials were poorly chosen and are not exploring the state space effectively. If

more than d^2 are large, then the system is experiencing strong non-Markovian effects (e.g., strong coupling to environmental degrees of freedom) or it has a larger Hilbert space dimension than expected.

	Fiducials	
#	Prep.	Measure
1		
2	Gx	Gx
3	Gy	Gy
4	Gx · Gx	Gx · Gx
5	Gx · Gx · Gx	Gx · Gx · Gx
6	Gy · Gy · Gy	Gy · Gy · Gy

#	Germ
1	Gx
2	Gy
3	Gi
4	Gx · Gy
5	Gx · Gy · Gi
6	Gx · Gi · Gy
7	Gx · Gi · Gi
8	Gy · Gi · Gi
9	Gx · Gx · Gi · Gy
10	Gx · Gy · Gy · Gi
11	Gx · Gx · Gy · Gx · Gy · Gy

Table 3: **Fiducial sequences and germs.** See discussion in text.

Quantity	Value
Number of strings	2737
Gate labels	Gx, Gy, Gi
SPAM labels	minus, plus
Counts per string	2000
Gram singular values (right column gives the values when using the target gate set)	0.0169 0 0.0309 0 0.8915 1 0.9042 1 0.911 1 2.9832 3

Table 4: **General dataset properties**. See discussion in text.

3 Output from GST

The primary output of GST is an estimated gate set. This section presents the raw estimate, and then some useful derived quantities of the estimated gates, including comparisons to the target gates.

3.1 Raw GST estimates

Table 5 reports the estimated SPAM operations, and Table 7 reports the logic gate operations. The estimated SPAM gates (ρ_0 and E_0) are vectors in $\mathcal{B}(\mathcal{H})$, and the estimated logic gates are superoperators represented as matrices acting on $\mathcal{B}(\mathcal{H})$, all in the Pauli basis. By taking the dot product of state preparation and measurement vectors, estimated SPAM probabilites are computed in Table 6.

The estimated gates can be compared directly to the target gate set given in Section 2. Ideally, they would match. In practice, of course, they won't. One of the best ways we have found to evaluate the significance of discrepancies is to compare *derived* quantities – i.e., certain properties calculated from the gate matrices and SPAM vectors. Deriving quantities from these raw outputs occupies the remainder of this section.

Operator	Hilbert-Schmidt vector (Pauli basis)	Matrix
ρ_0	0.7071 0.0003 0.0002 0.6091	$\begin{pmatrix} 0.9307 & 0.0003e^{-i0.620} \\ 0.0003e^{i0.620} & 0.0693 \end{pmatrix}$
E_0	0.7071 0.0004 0.0004 −0.7419	$\begin{pmatrix} -0.0246 & 0.0004e^{-i0.825} \\ 0.0004e^{i0.825} & 1.0246 \end{pmatrix}$

Table 5: **The GST estimate of the SPAM operations**. Compare to Table 1.

	E_0	E_C
ρ_0	0.0482	0.9518

Table 6: **GST estimate of SPAM probabilities**. Computed by taking the dot products of vectors in Table 5. The symbol E_C, when it appears, refers to the "complement" effect given by subtracting each of the other effects from the identity.

Gate	Superoperator (Pauli basis)
Gi	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -4\times10^{-5} & 0.997 & -0.0001 & -0.0001 \\ 2\times10^{-9} & 0.0001 & 0.997 & -4\times10^{-5} \\ -1\times10^{-5} & -3\times10^{-5} & -0.0001 & 0.997 \end{pmatrix}$
Gx	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 3\times10^{-5} & 0.9971 & 0.0001 & 3\times10^{-5} \\ 1\times10^{-5} & 0.0001 & -0.0001 & -0.9969 \\ 2\times10^{-5} & 6\times10^{-6} & 0.9969 & 0.0001 \end{pmatrix}$
Gy	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2\times10^{-5} & -4\times10^{-6} & -2\times10^{-5} & 0.997 \\ -1\times10^{-5} & -0.0001 & 0.997 & -0.0001 \\ -1\times10^{-5} & -0.997 & 0.0001 & -0.0001 \end{pmatrix}$

Table 7: **The GST estimate of the logic gate operations**. Compare to Table 2.

3.2 Derived quantities

Generally, the first thing that you want to know is "How far from ideal are the gates?". To answer this, the report tabulates several well-known definitions of distance. Table 8 lists the discrepancy from each estimated gate to its corresponding target, as measured by:

1. **Process infidelity**. Infidelity is simply $1 - F$, where F is the *fidelity*. The process fidelity between quantum processes G_a and G_b is given by $F = \mathrm{Tr}\left(\sqrt{\sqrt{\chi_a}\chi_b\sqrt{\chi_a}}\right)^2$, where χ_a and χ_b are the Jamiolkowski states (normalized Choi process matrices) corresponding to gate matrices G_a and G_b respectively. If the target is unitary (as is often the case), $F = \mathrm{Tr}\left(\chi_a\chi_b\right)$. Process infidelity is roughly what is measured in randomized benchmarking protocols; it quantifies the *incoherent* error rate if coherent errors (e.g. over-rotations) are not allowed to accumulate.

2. **Trace distance**. This is the *Jamiolkowski trace distance* between the Jamiolkowski states corresponding to the two processes: $d_{tr} = |\chi_a - \chi_b|_1 = \mathrm{Tr}\left(\sqrt{(\chi_a - \chi_b)^2}\right)$. This distance is useful primarily as a proxy for the *diamond norm distance*, because $d_{tr} \leq d_\diamond \leq \dim(\mathcal{H}) d_{tr}$.

3. **Diamond Norm**. The diamond norm between two quantum processes G_a and G_b is given by $\|G_a - G_b\|_\diamond = \sup_\rho \|(G_a \otimes I_k)(\rho) - (G_b \otimes I_k)(\rho)\|_1$, where I_k is the k-dimensional identity operation, $\|\cdot\|_1$ denotes the trace norm, and the supremum is taken over all $k \geq 1$ and density matrices ρ of dimension nk, with n the dimension of G_a and G_b. The diamond norm is also called the *completely bounded trace norm*, and plays the analogous role for quantum process distinguishability that the trace norm plays for density matrices. Specifically, the optimal probability of distinguishing G_a from G_b after a *single evaluation* is given by $\frac{1}{2} + \frac{1}{4}\|G_a - G_b\|_\diamond$. The diamond norm distance is an upper bound on the rate of error under any possible circumstance (including coherent accumulation of errors) and is often used in proofs of fault tolerance. For gates dominated by coherent/unitary error, it is common to see $d_\diamond \approx \sqrt{1 - F}$. For gates dominated by incoherent error, $d_\diamond \approx 1 - F$.

4. **Frobenius-norm distance**. The Frobenius norm distance between two gates G_a and G_b is simply $d_F = \sqrt{\mathrm{Tr}\left[(G_a - G_b)^2\right]}$. It has no known *operational* interpretation, but is very convenient as a rough measure of inaccuracy. It is also equal to the sum of the RMS errors in the individual matrix elements of the gates.

It's also useful to know *how* the real gates (or, more precisely, GST's estimates of the real gates) differ from the targets. There are several ways we could represent this, but the most useful involves an *error generator*. These are also given in Table 8. The final column of the table lists, for each gate, a Lindbladian superoperator $\mathbb{L}$. It is defined by the equation $\hat{G} = G_{\mathrm{target}} e^{\mathbb{L}}$, where $\hat{G}$ is the estimate and G_{target} is the ideal gate. This Lindbladian would be zero if the gates were perfect, and its overall magnitude is approximately equal to the diamond distance (or Jamiolkowski trace distance) between the target gate and the estimate.

It's usually useful to understand *how* gates fail. The error generators in Table 8 provide one view on this, but they are not necessarily intuitive. For example, you might want to know whether your gate suffers depolarizing, dephasing, or over-rotation errors. In Table 9, the estimated gates are decomposed into: (1) rotations (including angle and axis errors); (2) incoherent *diagonal* decay rates (depolarizing or T_1 noise); and (3) incoherent *off-diagonal* decay rates (dephasing or T_2 noise). These analyses can be compared with a the similar decomposition of the target gates (cf. table 2). Note that for some erroneous gates, this decomposition simply fails; if the numbers make no sense, this is probably the case.

It might be useful to know the closest *unitary* operation to the estimated gate, and how close it is. Usually, you were trying to implement a unitary. If the closest unitary to G was indeed G_{target}, then all errors are incoherent; if not, you might be able to tweak the gate parameters to get closer relatively easily. Also, implementing a particular unitary may be less important than just achieving *some* set of mutually independent unitaries. In these and other cases, the distance from an estimated gate to its closest unitary approximation is of interest.

Table 10 lists, for each estimated gate, the properties of its closest unitary approximation. The table defines the closest unitary, in terms of an axis and angle (in $\mathcal{B}(\mathcal{H})$) of rotation. It also presents the process fidelity and Jamiolkowski trace distance between the estimated gate and its closest unitary approximation.

Gate	Process Infidelity	1/2 Trace Distance	1/2 ◇-Norm
Gi	0.0022	0.0022	0.0022
Gx	0.0023	0.0023	0.0023
Gy	0.0023	0.0023	0.0023

Gate	Error Generator
Gi	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ -4\times10^{-5} & -0.003 & -0.0001 & -0.0001 \\ 4\times10^{-9} & 0.0001 & -0.003 & -4\times10^{-5} \\ -1\times10^{-5} & -3\times10^{-5} & -0.0001 & -0.003 \end{pmatrix}$
Gx	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 3\times10^{-5} & -0.0029 & 0.0001 & 3\times10^{-5} \\ 2\times10^{-5} & 6\times10^{-6} & -0.0031 & 0.0001 \\ -1\times10^{-5} & -0.0001 & 0.0001 & -0.0031 \end{pmatrix}$
Gy	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1\times10^{-5} & -0.003 & -0.0001 & 0.0001 \\ -1\times10^{-5} & -0.0001 & -0.003 & -0.0001 \\ 2\times10^{-5} & -4\times10^{-6} & -2\times10^{-5} & -0.003 \end{pmatrix}$

Table 8: **Comparison of GST estimated gates to target gates**. This table presents, for each of the gates, three different measures of distance or discrepancy from the GST estimate to the ideal target operation. See text for more detail. The second table lists the "Error Generator" for each gate, which is the Lindbladian $\mathbb{L}$ that describes *how* the gate is failing to match the target. This error generator is defined by the equation $\hat{G} = G_{\text{target}}e^{\mathbb{L}}$.

A sanity check is computed by comparing the fidelity of the obtained closest unitary with a theoretical upper bound (if a value greater than one appears in this column then the other values in that row may be inaccurate). If these numbers are similar to those in Table 8, then the gates are as close to the targets as they are to *any* unitary.

Finally, Table 11 presents each estimated gate's *Choi matrix*, along with its spectrum. The Choi matrix (sometimes ambiguously referred to as the "process matrix") is an alternative way to describe a process. We usually prefer the "superoperator representation", which has the very useful property that the process matrix corresponding to applying G_a and then G_b is simply G_bG_a. This is completely false for the Choi representation. Nonetheless, the Choi representation is often useful, so we present it here – but without a detailed discussion of its properties (see, e.g. the textbook by Nielsen and Chuang).

The Choi matrix $\chi(G)$ for a gate G can be simply understood in either of two ways. First, it is equivalent (up to choice of basis) to the *Jamiolkowski state* defined by applying G to one half of a maximally entangled bipartite state. Second, it is the general (non-diagonal) form of the well-known Kraus representation, $G[\rho] = \sum_i K_i \rho K_i^\dagger$. The Choi matrix behaves in many ways like a quantum state, and appears naturally in expressions for the process fidelity and Jamiolkowsi trace distance just as density matrices would enter these expressions when computing differences between states.

Additionally, the condition of *complete positivity* or CP (which all real quantum processes must satisfy) is simply the positivity of the Choi matrix. Thus, negative eigenvalues in Table 11 indicate that the estimate violates complete positivity. If they are very small, they may simply indicate statistical fluctuations (unitary gates have χ matrices with zero eigenvalues, so any small fluctuation is likely to violate CP). If they are large, they serve as a warning that (1) the model of CPTP maps is probably violated (usually because of non-Markovian behavior), and (2) this estimate may produce negative or greater-than-unity probabilities. GST does *not* generally impose complete positivity (although it is an option), precisely because violation of CP is a warning flag for non-Markovian behavior (which is very common in experimental qubits).

Gate	Eigenvalues	Fixed pt	Rotn. axis	Diag. decay	Off-diag. decay
Gi	0.9971 $0.997e^{i0.000}$ $0.997e^{-i0.000}$ 1	0.9998 −0.0146 −0.0005 −0.0039	0 −0.278 −0.7336 0.6201	0.0029	0.003
Gx	$0.9969e^{i1.571}$ $0.9969e^{-i1.571}$ 0.9971 1	0.9999 0.0088 -1×10^{-6} 1×10^{-5}	0 1 2×10^{-5} 3×10^{-5}	0.0029	0.0031
Gy	$0.997e^{i1.571}$ $0.997e^{-i1.571}$ 0.997 1	1 3×10^{-6} −0.0035 -1×10^{-5}	0 0.0001 1 0.0001	0.003	0.003

Gate	Angle	Angle between Rotation Axes		
		Gi	Gx	Gy
Gi	$3\times10^{-5}\pi$		–	–
Gx	0.5π	–		0.5π
Gy	0.5π	–	0.5π	

Table 9: **Eigen-decomposition of estimated gates**. Each estimated gate is described in terms of: (1) the eigenvalues of the superoperator; (2) the gate's fixed point (as a vector in $\mathcal{B}(\mathcal{H})$, in the Pauli basis); (3) the axis around which it rotates, as a vector in $\mathcal{B}(\mathcal{H})$; (4) the angle of the rotation that it applies; (5) the decay rate along the axis of rotation ("diagonal decay"); (6) the decay rate perpendicular to the axis of rotation ("off-diagonal decay"); and (7) the angle between each gate's rotation axis and the rotation axes of the other gates. "X" indicates that the decomposition failed or couldn't be interpreted.

Gate	Process Infidelity	1/2 Trace Distance	Rotation Axis	Rotation Angle	Sanity ✓
Gi	0.0022	0.0022	0 −0.1419 −0.4529 0.8802	$3\times10^{-5}\pi$	6×10^{-8}
Gx	0.0023	0.0023	0 −1 -4×10^{-5} -1×10^{-6}	0.5π	3×10^{-8}
Gy	0.0023	0.0023	0 4×10^{-5} 1 -2×10^{-6}	0.5π	1×10^{-8}

Table 10: Information pertaining to the closest unitary gate to each of the estimated gates.

Gate	Choi matrix (Pauli basis)				Eigenvalues
Gi	0.9978	$1 \times 10^{-5} e^{i2.554}$	$-2 \times 10^{-5} i$	$5 \times 10^{-5} e^{-i1.638}$	0.0007
	$1 \times 10^{-5} e^{-i2.554}$	0.0008	$8 \times 10^{-6} e^{i2.769}$	-4×10^{-5}	0.0007
	$2 \times 10^{-5} i$	$8 \times 10^{-6} e^{-i2.769}$	0.0008	$3 \times 10^{-5} e^{i0.370}$	0.0008
	$5 \times 10^{-5} e^{i1.638}$	-4×10^{-5}	$3 \times 10^{-5} e^{-i0.370}$	0.0007	0.9978
Gx	0.4993	$0.4985 e^{-i1.571}$	$7 \times 10^{-6} e^{i2.058}$	$9 \times 10^{-6} e^{i1.097}$	0.0007
	$0.4985 e^{i1.571}$	0.4993	$3 \times 10^{-5} e^{-i3.027}$	$1 \times 10^{-5} e^{-i0.337}$	0.0008
	$7 \times 10^{-6} e^{-i2.058}$	$3 \times 10^{-5} e^{i3.027}$	0.0007	$6 \times 10^{-6} e^{-i1.512}$	0.0008
	$9 \times 10^{-6} e^{-i1.097}$	$1 \times 10^{-5} e^{i0.337}$	$6 \times 10^{-6} e^{i1.512}$	0.0008	0.9977
Gy	0.4992	$0.0001 e^{-i1.500}$	$0.4985 e^{i1.571}$	$1 \times 10^{-5} e^{i1.864}$	0.0007
	$0.0001 e^{i1.500}$	0.0008	$2 \times 10^{-5} e^{i0.136}$	$3 \times 10^{-6} e^{i1.447}$	0.0007
	$0.4985 e^{-i1.571}$	$2 \times 10^{-5} e^{-i0.136}$	0.4993	$9 \times 10^{-6} e^{-i2.634}$	0.0008
	$1 \times 10^{-5} e^{-i1.864}$	$3 \times 10^{-6} e^{-i1.447}$	$9 \times 10^{-6} e^{i2.634}$	0.0007	0.9977

Table 11: **Choi matrix representation of the GST estimated gate set**. This table lists Choi representations of the estimated gates, and their eigenvalues. Unitary gates have a spectrum $(1, 0, 0 \ldots)$, just like pure quantum states. Negative eigenvalues are non-physical, and may represent either statistical fluctuations or violations of the CPTP model used by GST.

4 Goodness-of-model Analysis

The previous section presented the estimated gate set, and compared it to the target gate set. This section is concerned with a mostly orthogonal analysis which seeks to explain how much the estimated gate set can be trusted – i.e., how well it fits the data.

To understand the goal of this section, consider the simple problem of fitting a line to a set of points. For any set of points, there is *always* a best-fit line – but this doesn't mean that the best-fit line is a *good* fit! The data points may trace out a parabola, a square, or even something more complicated. It is essential to understand not just what the best-fit line was (and perhaps how close it was to some desired line), but also **how well that linear model was able to fit all the data**. Of course, we do not expect it to fit every data point perfectly. The critical question is "Did the linear model fit *as well as we would expect it to* if the data really were generated by a linear process?"

In this analogy, GST's estimated gate set is like the best-fit line, and the target gate set like the desired line. This section asks the question "How well was GST able to fit all of the data – and did it fit well enough to suggest that its model is valid?". A central tool used to do this is the *likelihood function*, which we denote $\mathcal{L}$, which formally is the probability of the observed data given a set of model parameters. The basic idea is that we maximize the likelihood function to obtain the best set of model parameters (i.e. gate set), and by looking at the value of this maximum we can determine the model's goodness-of-fit. We will actually deal primarily with the logarithm of the likelihood function, $\log(\mathcal{L})$, which is simliarly maximized.

4.1 Aggregated $\log(\mathcal{L})$

The log-likelihood for an n-outcome system with predicted probabilities p_i and observed frequencies f_i $(i = 1 \dots n)$ is given by:

$$\log(\mathcal{L}) = \sum_i N f_i \log(p_i). \tag{1}$$

where N is the total number of counts. In *this* analysis, $\log(\mathcal{L})$ is used to compare the set of probabilities predicted by a gate set (p_s) and the frequencies obtained from a dataset (f_s). Each experiment (or gate sequence) s is associated to two probabilities: "plus" has probability p_s and "minus" has probability $1 - p_s$. The $\log(\mathcal{L})$ contribution of a single gate string s is

$$\log(\mathcal{L})_s = N f_s \log(p_s) + N(1 - f_s) \log(1 - p_s), \tag{2}$$

where N is the number of times the experiment s was performed, p_s is the probability of a "plus" outcome as predicted by the gate set, and f_s is the observed frequency of "plus". The total log-likelihood for an entire dataset is just the sum

$$\log(\mathcal{L}) = \sum_{s \in \mathcal{S}} \log(\mathcal{L})_s. \tag{3}$$

A theoretical upper bound on the log-likelihood can be found by replacing p_s with f_s in Eq. 2 and evaluating Eq. 3. We will refer to this quantity as $\log(\mathcal{L})_{ub}$.

Statistical theory has quite a lot to say about the likelihood function (see any of the major textbooks). Using some of these results, we can predict that if there are N_p free parameters in the gate set that GST is fitting, and GST fits a dataset containing $N_s > N_p$ distinct experiments (gate sequences), then *if the gate set model is correct*, then two times the difference between $\log(\mathcal{L})_{ub}$ and the maximum $\log(\mathcal{L})$ obtained is a random variable with a χ^2_k distribution, where

$$k \equiv N_s - N_p.$$

Its expected value is $\langle \chi^2 \rangle = k$, and its standard deviation is $\sqrt{2k}$. Thus, if the fit is "good", then twice $\Delta \log(\mathcal{L}) \equiv \log(\mathcal{L})_{ub} - \max(\log(\mathcal{L}))$ should lie roughly within the interval $[k - \sqrt{2k}, k + \sqrt{2k}]$. Thus, by comparing the difference $2\Delta \log(\mathcal{L}) - k$ to $\sqrt{2k}$, one can determine how well the GST estimate was able to fit the data in dataset "test".

The MLEGST algorithm used to generate this estimate is iterative. It starts by fitting only data from the shortest gate sequences (which are easy to fit *and* insensitive to most non-Markovian noise), then successively

adds longer and longer sequences (with base sequence length $L \leq 1, 2, 4, 8, \ldots$) to the mix. Since we get an estimate at each intermediate L, it is possible to quantify not just the goodness of the *best* fit (presented in the previous section), but how the goodness-of-fit behaves as longer and longer sequences are added in.

This data is presented in Table 12. What you should be looking for here is whether – at each value of L – the $2\Delta\log(\mathcal{L})$ quantity is roughly the same as k. More precisely, is $|2\Delta\log(\mathcal{L}) - k|$ less than or equal to $\sqrt{2k}$? If not, then the model is not fitting as well as it should, which usually indicates non-Markovian noise (or, rarely, that the GST algorithm has simply failed to find a good fit even though one exists).

As a rough rule of thumb, for GST experiments involving relatively long sequences (e.g. $L \geq 100$):

- "Incredibly good" (★★★★★) experiments have $2\Delta\log(\mathcal{L}) \approx k$, as predicted by theory (and seen in simulations).
- "Great" (★★★★) experiments have $2\Delta\log(\mathcal{L}) \leq 2k$ or so.
- "Good" (★★★) experiments have $2\Delta\log(\mathcal{L}) \leq 5k$ or so.
- "Okay" (★★) experiments have $2\Delta\log(\mathcal{L}) \leq 10k$.
- Experiments in which $2\Delta\log(\mathcal{L}) > 10k$ (★) have very significant non-Markovian noise, and the results in the previous section should be viewed very cautiously.

L	$2\Delta\log(\mathcal{L})$	k	$2\Delta\log(\mathcal{L}) - k$	$\sqrt{2k}$	p	N_s	N_p	Rating
0	65.664	61	4.6635	11.045	0.32	92	31	★★★★★
1	65.664	61	4.6638	11.045	0.32	92	31	★★★★★
2	159.42	137	22.419	16.553	0.09	168	31	★★★★
4	476.3	410	66.299	28.636	0.01	441	31	★★★★
8	903.82	786	117.82	39.648	2×10^{-3}	817	31	★★★★
16	1274.2	1170	104.16	48.374	0.02	1201	31	★★★★
32	1657.5	1554	103.53	55.749	0.03	1585	31	★★★★
64	2062.1	1938	124.08	62.258	0.02	1969	31	★★★★
128	2490.4	2322	168.4	68.147	0.01	2353	31	★★★★
256	2831.2	2706	125.2	73.566	0.05	2737	31	★★★★

Table 12: **Comparison between the computed and expected maximum** $\log(\mathcal{L})$ **for different values of** L. N_S and N_p are the number of gate strings and parameters, respectively. The quantity $2\Delta\log(\mathcal{L})$ measures the goodness of fit of the GST model (small is better) and is expected to lie within $[k-\sqrt{2k}, k+\sqrt{2k}]$ where $k = N_s - N_p$. p is the p-value derived from a χ^2_k distribution.(For example, if $p = 0.05$, then the probability of observing a χ^2 value as large as, or larger than, the one indicated in the table is 5%, assuming the GST model is valid.) The rating from 1 to 5 stars gives a very crude indication of goodness of fit as explained in the text.

4.2 Detailed likelihood analysis

The aggregated $2\Delta\log(\mathcal{L})$ numbers presented in Table 12 tell you how well the GST estimate fits the *entire* dataset. If they are in line with theory ($2\Delta\log(\mathcal{L}) \approx k$), then there is little more to be said. But if the best fit to the data is not good, we can debug it by identifying *which* experiments are inconsistent with the fit.

Figure 1 displays the $2\Delta\log(\mathcal{L})$ contribution from each individual gate sequence (Eq. 2). Each gate sequence corresponds to a single colored "pixel" in the plot. Each block of pixels corresponds to a single base sequence (i.e., a germ power), and the individual pixels within a block correspond to the various fiducial sequence pairs between which that base sequence was sandwiched. (The column indicates the fiducial adjacent to state preparation, while the row indicates the fiducial adjacent to measurement). Base sequences are arranged in a grid; different rows correspond to different germs, while different columns correspond to different maximum lengths L. Pixels are labeled with the $2\Delta\log(\mathcal{L})$ contribution for that sequence, and colored appropriately.

Sequences whose observed frequencies are consistent with a Markovian gate set are shown in gray, with darker shades indicating greater inconsistency with the estimated gate set. Data shown in red are *not* consistent with a Markovian gate set. It may appear contradictory to say that (a) gray is "consistent" with Markovian, but (b) darker shades indicate "greater inconsistency". The resolution is that the χ^2 values quantify inconsistency with the model, *but* they themselves are also subject to random fluctuations. Therefore, even if the data are perfectly consistent with the model, we expect to see (for example) a single $\chi^2_s \geq 10$ once per each 638 experiments. Observing $\chi^2_s \geq 10$ for any given sequence does suggest that the data from s were relatively surprising, but we also expect to see one such fluctuation if there are more than about 600 experiments. The gray/red threshold is chosen based on the total number of sequences so that *if* the data are perfectly Markovian, then the probability of one or more experiments being colored red is only 5%.

Identifying patterns and trends within such "pixel plots" can aid in identifying specific sources and types of non-Markovian noise which may be to blame if the GST algorithms are unable to produce a "good" estimate. For example, it is often the case that all the short sequences [$L = O(1)$] can be fit reasonably well, but the right-hand side of Figure 1 becomes a sea of red. This indicates that non-Markovian behavior (potentially due to slow drift of gate set parameters) is becoming more significant for longer experiments. In other cases, a single row may be particularly bad, indicating that a particular gate or germ is especially problematic (e.g., was not stabilized using dynamical decoupling techniques). Be cautious in debugging, however – sometimes bad $\log(\mathcal{L})$ values for a particular gate or germ can result *not* from faults in that operation, but because another operation failed so badly that it distorted the entire fit (e.g., in trying to fit catastrophically non-Markovian data at Point A, GST ended up failing to fit perfectly good data at Point B).

Similar pixel plots for the intermediate estimates whose total $2\Delta\log(\mathcal{L})$ is listed in Table 12 are included when the "pixel plots" appendix is enabled.

Similar pixel plots for the intermediate estimates whose total $2\Delta\log(\mathcal{L})$ is listed in Table 12 can be found in Appendix

Figure 2 shows exactly the same $\log(\mathcal{L})$ analysis as Figure 1, but arranged differently. Here, blocks (not square) all correspond to a single fiducial pair (e.g., pre- and post-fiducial), and pixels within a block correspond to different base sequences. This can be useful for diagnosing a single bad fiducial sequence.

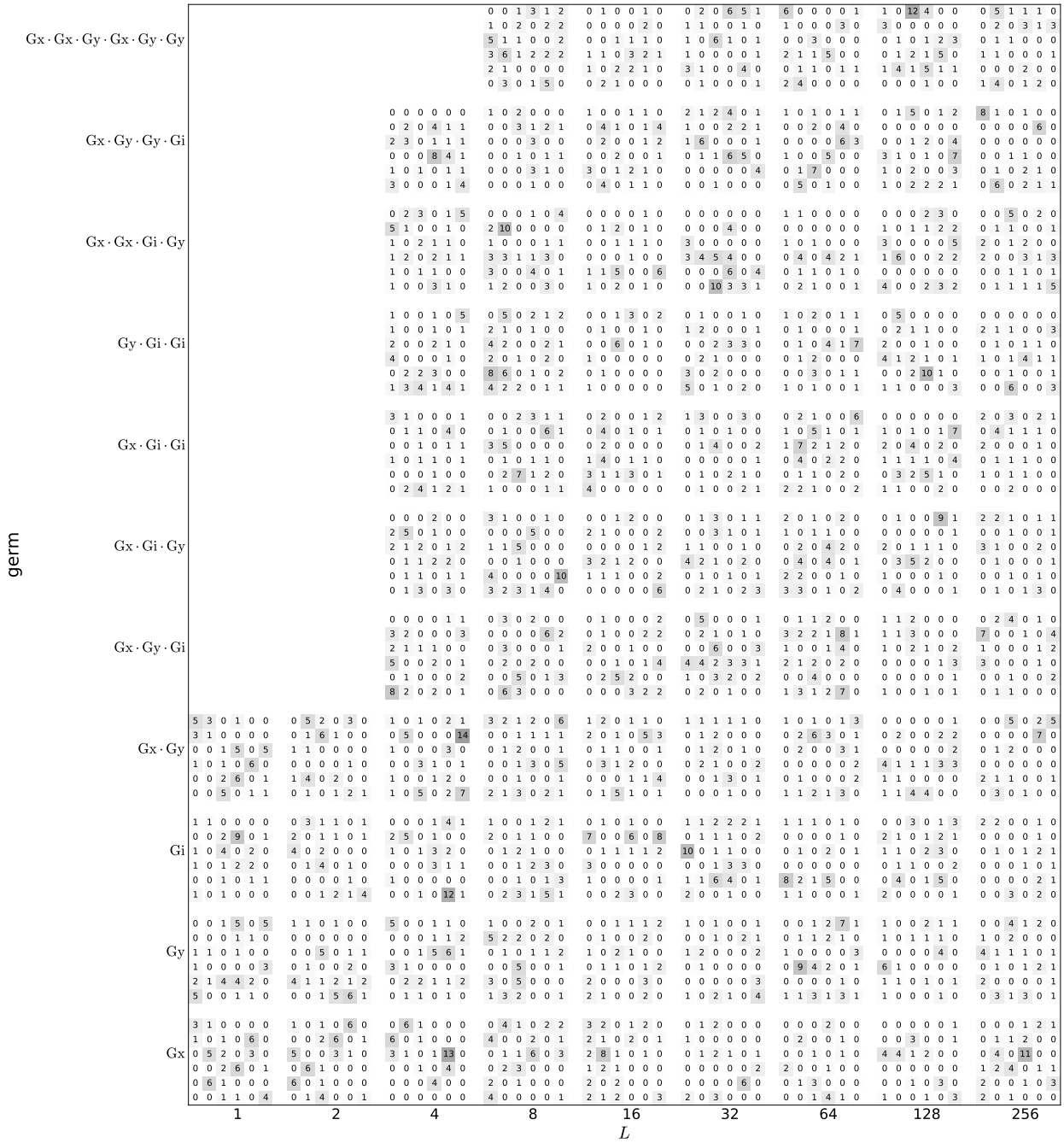

Figure 1: $2\Delta\log(\mathcal{L})$ **contributions for every individual experiment in the dataset**. Each pixel represents a single experiment (gate sequence), and its color indicates whether GST was able to fit the corresponding frequency well. Shades of white/gray are typical. Red squares represent statistically significant evidence for model violation (non-Markovianity), and should appear with probability at most 5% if the data really are Markovian. Square blocks of pixels correspond to base sequences (arranged vertically by germ and horizontally by length); each pixel within a block corresponds to a specific choice of pre- and post-fiducial sequences. See text for further details.

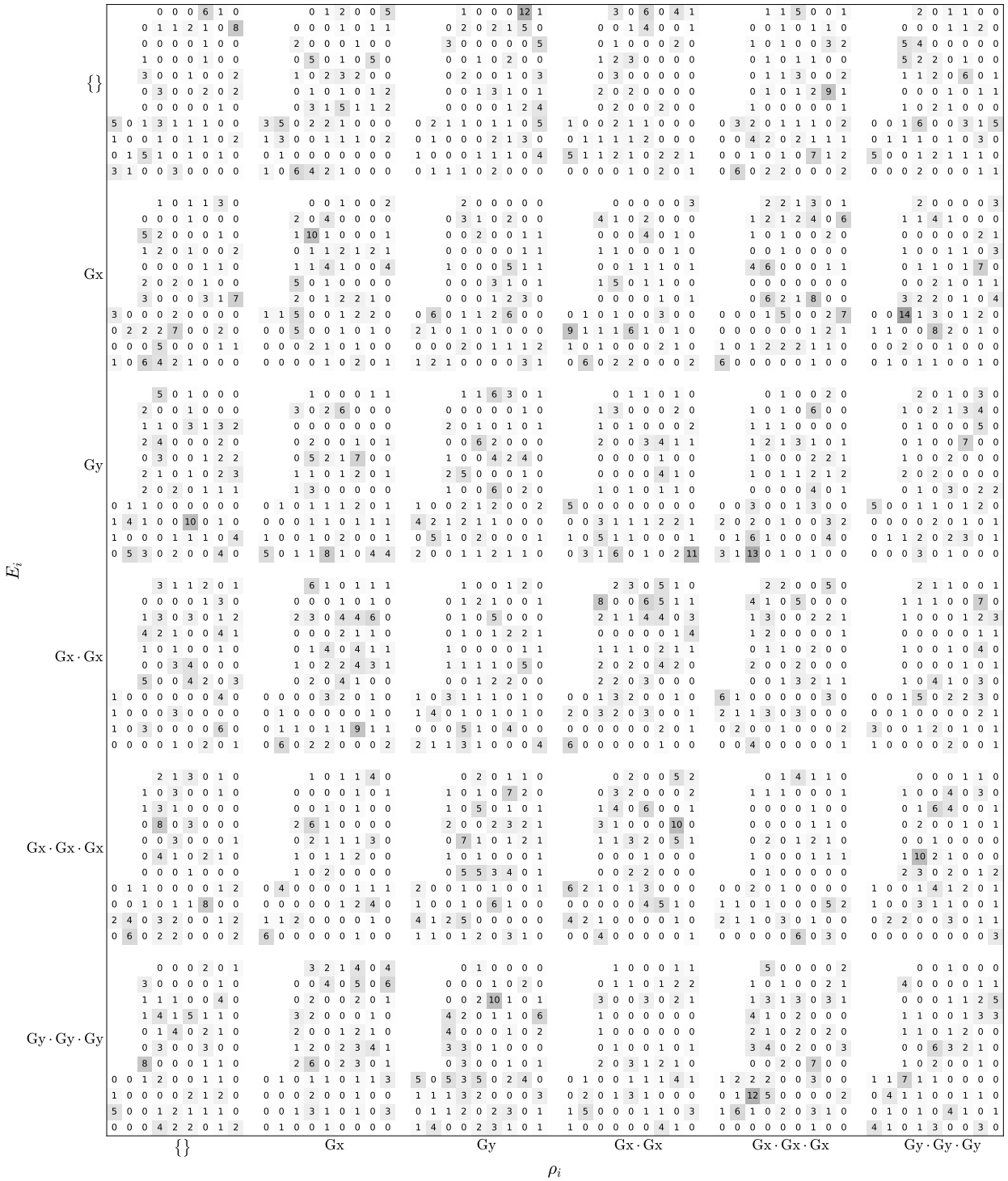

Figure 2: $2\Delta\log(\mathcal{L})$ **contributions for each experiment, arranged differently.** This figure shows the same data as Figure 1, but arranged differently. Each block now corresponds to a particular pair of fiducial sequences, while pixels within the block correspond to different base sequences sandwiched between those fiducials.