New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG IOP Institute of Physics

Accelerated randomized benchmarking

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

CrossMark

OPEN ACCESS

RECEIVED
30 September 2014

ACCEPTED FOR PUBLICATION
8 December 2014

PUBLISHED
20 January 2015

PAPER

Accelerated randomized benchmarking

Christopher Granade[1,2], **Christopher Ferrie**[3] **and D G Cory**[1,4,5,6]

[1] Institute for Quantum Computing, University of Waterloo, Waterloo, ON, Canada
[2] Department of Physics, University of Waterloo, Waterloo, ON, Canada
[3] Center for Quantum Information and Control, University of New Mexico, Albuquerque, NM 87131-0001, USA
[4] Department of Chemistry, University of Waterloo, Waterloo, ON, Canada
[5] Perimeter Institute, Waterloo, ON, Canada
[6] Canadian Institute for Advanced Research, Toronto, ON, Canada

E-mail: cgranade@cgranade.com

Keywords: quantum information, randomized benchmarking, characterization

Abstract

Quantum information processing offers promising advances for a wide range of fields and applications, provided that we can efficiently assess the performance of the control applied in candidate systems. That is, we must be able to determine whether we have implemented a desired gate, and refine accordingly. Randomized benchmarking reduces the difficulty of this task by exploiting symmetries in quantum operations. Here, we bound the resources required for benchmarking and show that, with prior information, we can achieve several orders of magnitude better accuracy than in traditional approaches to benchmarking. Moreover, by building on state-of-the-art classical algorithms, we reach these accuracies with near-optimal resources. Our approach requires an order of magnitude less data to achieve the same accuracies and to provide online estimates of the errors in the reported fidelities. We also show that our approach is useful for physical devices by comparing to simulations.

Quantum information processing devices offer great promise in a variety of different fields, including chemistry and material science, data analysis and machine learning [1–4], as well as cryptography [5]. Over the past few years, proposals have been advanced for quantum information processing past the classical scale, based on node-based architectures [6, 7]. In addition, rapid progress has been made towards experimental implementations that might allow for developing such devices [8, 9]. An impediment in this effort, however, is presented by the difficulty of calibrating and diagnosing quantum devices.

In particular, in the development of quantum information processing, an important experimental challenge is to efficiently characterize the quality with which we can control a quantum system. By characterizing the quality of a quantum gate that is implemented by a control pulse, we can then reason about the utility of that gate for quantum information processing tasks. For instance, we can estimate the feasibility of and the resources required to implement error correction using that control by comparing to proven and numerically estimated fault-tolerance thresholds [10, 11]. Alternately, we can adjust our control sequences to account for differences between our control model and the actual system.

In cases where only the quality of a quantum gate or set of gates is required, randomized benchmarking has proven to be a useful means of extracting this information with relatively little experimental effort [12]. This has been demonstrated in a variety of experimental settings [9, 13–19]. Randomized benchmarking has also been used to improve gate fidelities by characterizing cross-talk [20] or distortions [21]. Extracting fidelity information can often be useful in diagnosing performance and problems with a device in lieu of full characterization [22]. Moreover, randomized benchmarking has also been used to extract information about the completely positive and unital parts of linear maps [23].

Here, using near-optimal data processing together with prior information, we accelerate the data processing used in benchmarking experiments, such that to achieve the accuracy demanded of benchmarking protocols, we require orders of magnitude less experimental data. We also extend results on the achievable estimation quality in the presence of finite sampling [24] and prior information, then show that our accelerated methods are nearly optimal. Our data processing methods also provide estimates of their own performance, such that our approach thus enables randomized benchmarking to be used where data collection costs make existing benchmarking protocols impractical. Thus, our work complements recent results on the robustness of randomized benchmarking [25] to provide an experimentally useful tool[7].

Randomized benchmarking has been recently used to adaptively calibrate control designed by optimal control theory methods such as GRAPE [26], allowing for differences between the control model and the actual system to be adjusted for in experimental practice [27]. These methods are applied in a control design and calibration step, however, and do not allow for control for to be recalibrated *dynamically*. Whereas randomized benchmarking is performed at the inner-loop of current control calibration algorithms [22], any data collection overhead in benchmarking becomes a very significant cost to control calibration as a whole. Thus, by reducing the data requirements using both better fitting methods and strong prior information, we can enable new applications, such as extending control calibration to an online context.

Here, we show that by using prior information together with the sequential Monte Carlo (SMC) parameter estimation algorithm, we can obtain very accurate estimates of parameters. Moreover, we do so even in the limit of one bit of data per sequence length. That is, we can use a variety of sequence lengths to probe the performance of our gate set rather than repeat many experiments at a given sequence length. On the other hand, we also show that for gates with fidelities near unity, increasing the length of benchmarking sequences offers little compared to repeating experiments at already optimal sequence lengths. The SMC algorithm is based on Bayesian methods, which have been used successfully in a variety of quantum information processing tasks [28–34]. SMC has recently been used in quantum information to learn states [30] and Hamiltonians [35, 36], and to provide robust error bounds on inferred parameters [37]. The primary cost incurred by the SMC algorithm is that the data must be simulated repeatedly, though this can be mitigated by using quantum resources [38–40]. Here we show that since the symmetries afforded by random benchmarking experiments can be used to simulate datasets with costs that are constant with respect to the dimension of the Hilbert space of interest [12], SMC can be implemented with little overhead. Thus, randomized benchmarking mitigates the primary disadvantage of SMC by removing the need to simulate the quantum system.

Moreover, the method of hyperparameters [35] generalizes our approach to allow gate fidelities to be non-trivial functions of some other parameter of interest, such that the underlying parameter is learned directly. This approach is especially relevant if, for example, the effect of the unknown hyperparameter depends on an experimental choice, such that distinct benchmarking experiments can be used in concert in a straightforward way.

Our work proceeds first by defining the benchmarking model that we use, then showing bounds on the estimation of the parameters of this model using the Cramer–Rao bound. We then apply SMC to the benchmarking model and compare to the performance of traditional methods, and to the optimal performance achievable with prior information, showing that our method offers distinct advantages, and is nearly optimal.

1. Interpretation of likelihood as marginalization

Randomized benchmarking consists of using a sequence of random gates to effectively average the action of an error channel such that it can be simulated using simple classical models. If the gates in each randomized benchmarking sequence are chosen uniformly at random from the Clifford group, then the argument of Magesan *et al* [12] shows that the average fidelity F_g taken over all randomized benchmarking sequences of a given length can be expressed in terms of the *survival probability*

$$\Pr\left(\text{survival}|\psi,\, \boldsymbol{i}_m\right) = \operatorname{Tr}\left[E_\psi \hat{\hat{S}}_{\boldsymbol{i}_m}(\rho_\psi)\right], \tag{1}$$

where E_ψ is a measurement operator corresponding to a fiducial state ρ_ψ , and where $\hat{\hat{S}}_{\boldsymbol{i}_m} = \hat{\hat{S}}_{i_m} \circ \cdots \circ \hat{\hat{S}}_{i_1}$ is the superoperator representing the sequence $\boldsymbol{i}_m$. Because the Clifford group C forms a unitary two-design, random sequences of Clifford gates average the errors in each gate over the Haar measure, an operation known as *twirling*. In particular, given a channel Λ, conjugating the action of that channel by ideal Clifford gates chosen uniformly at random implements the twirling superchannel [12],

$$\begin{aligned} W[\Lambda](\rho) &= \int \mathrm{d}U\; U^\dagger \Lambda\left[U\rho U^\dagger\right]U \\ &= \frac{1}{|C|}\sum_{U\in C} U^\dagger\Lambda\left[U\rho U^\dagger\right]U = p\rho + (1-p)\frac{\mathbb{1}}{d}, \end{aligned} \tag{2}$$

[7] Literate source code for this work is available at https://github.com/cgranade/accelerated-randomized-benchmarking. To view the code online, visit http://nbviewer.ipython.org/github/cgranade/accelerated-randomized-benchmarking/blob/master/src/model_testing.ipynb.

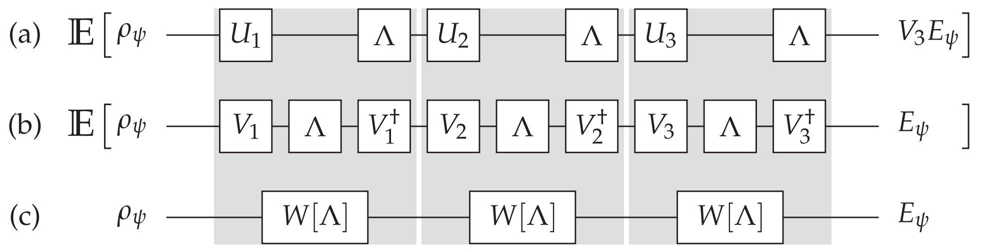

Figure 1. Sketch of Magesan *et al* derivation of the zeroth-order model [12]. (a) Sequence of length $m = 3$ Clifford operations. (b) Change of variables to V_i, factoring out previous gates U_{i-1}, and with the base case $U_1 = V_1$. The V gates then form a two-design. (c) Expectation value over random gates in (a) and (b), giving the twirling superchannel W acting on Λ.

where d is the dimension of the Hilbert space on which each gate acts, and where p is related to the average gate fidelity F of Λ by $p = (dF - 1)/(d - 1)$.

The expectation value of this survival probability over all sequences of a given length m was shown to produce the uniform-average fidelity

$$F_g(m, \psi) = \mathbb{E}_{\boldsymbol{i}_m|m}\left[\Pr\left(\text{survival}|\psi, \boldsymbol{i}_m\right)\right] = \Pr(\text{survival}|\psi, m). \tag{3}$$

We may thus interpret the fidelity averaged over a unitary design as a probability of survival in an experiment in which we do not know the sequence being performed. As discussed in detail in appendix, if sequences are fairly drawn from the two-design independently of all other experimental choices, then this is a valid assumption, such that the marginalized survival probability can be taken as the likelihood for our randomized benchmarking model. Note that in the remainder of the paper, we will let ψ be fixed, and will drop the notation conditioning on this assumption.

Using the expansion of the marginalized survival $F_g(m)$ given by Magesan *et al* [12], we can rewrite the likelihood in a way that explicitly depends on the parameters of interest, and that no longer requires simulating the quantum dynamics of the system. Thus, we can use Bayesian methods without simulating the system under study. In [12], the authors studied a sequence of models of increasing complexity, where the lowest complexity model has found the most use. In particular, we consider the Magesan *et al* model [12], which they call the *zeroth-order model* for randomized benchmarking

$$F_g(m) = A_0 p^m + B_0 \tag{4}$$

for parameters A_0, B_0 which encode errors in preparation and measurement, and p. These parameters are given formally by

$$A_0 := \mathrm{Tr}\left[E_\psi \Lambda\left(\rho_\psi - \frac{\mathbb{1}}{d}\right)\right] \tag{5a}$$

$$B_0 := \mathrm{Tr}\left[E_\psi \Lambda\left(\frac{\mathbb{1}}{d}\right)\right] \tag{5b}$$

$$p := \left(dF_{\text{ave}} - 1\right)/(d - 1), \tag{5c}$$

where d is the dimension of the Hilbert space. Above, F_{ave} is the fidelity of the average channel $\Lambda = \mathbb{E}_{i,j}[\Lambda_{i,j}]$, taken over time steps i and elements of the gate set j. By these definitions, for ideal preparation, evolution and measurement, $A_0 = 1 - \frac{1}{d}$ and $B_0 = \frac{1}{d}$. Since we will often use the example of a qubit, we thus have that the ideal $A_0 = B_0 = 1/2$. A sketch of the derivation of this model is given in figure 1. The interpretation of first- and higher-order models follows in a similar manner. Since we use the zeroth-order model as an example in this work, we will drop the subscript-0 for brevity.

Because the fidelity of a channel is invariant under Clifford twirling, the parameter p represents the strength of the depolarizing channel of fidelity F_{ave} produced by twirling the average channel Λ, and can be used to recover F_{ave}. Similarly, in the interleaved protocol [41], we consider two probabilities, p_{ref} and $p_{\bar{\mathcal{C}}}$, respectively representing the sequences with m random Clifford gates multiplied together, or interleaved with some gate $\mathcal{C}$ under study. From these probabilities, we can extract the referenced probability of gate error $\tilde{p} := p_{\text{ref}}/p_{\bar{\mathcal{C}}}$. Each of p_{ref} and $p_{\bar{\mathcal{C}}}$ is traditionally extracted from a fit to the zeroth- or first-order model[8].

[8] We note that following the central limit theorem, the estimators $\hat{p}_{\text{ref}}$ and $\hat{p}_{\bar{\mathcal{C}}}$ will be approximately normally distributed about the true values of each parameter. Thus, estimating $\tilde{p}$ from $\hat{p}_{\text{ref}}/\hat{p}_{\bar{\mathcal{C}}}$ results in an estimator that is Cauchy-distributed and therefore has no defined mean or variance. Estimates of the bias or error for this procedure therefore cannot be robustly provided by considering the sample standard deviations reported by least-squares fitting software.

2. Achievable accuracy

We now consider only the interleaved model since it is more general. For brevity, we represent the model by a vector $\boldsymbol{x} = (\tilde{p},\ p_{\text{ref}},\ A,\ B)$, so that the likelihood function for the interleaved model is

$$\Pr(1|\boldsymbol{x};\ m,\ \text{mode}) = \begin{cases} Ap_{\text{ref}}^{m} + B & \text{mode is reference,} \\ A\left(p_{\text{ref}}\tilde{p}\right)^{m} + B & \text{mode is interleaved,} \end{cases} \tag{6}$$

where we have labeled the survival event by '1' to more easily allow for using binomial distributions to consider sums over multiple measurements of the same sequence length.

Having defined our model, it is critical to account for the accuracy with which we can estimate the parameters using finite data records. Here, we extend the results of Epstein *et al* [24] by explicitly calculating the Fisher information of $\Pr(1|\boldsymbol{x})$. We can find a bound on the achievable estimation error in this model by appealing to the Cramér–Rao bound [42], which states that the Fisher information matrix $\boldsymbol{I}(\boldsymbol{x})$ bounds the error matrix $\boldsymbol{E}(\boldsymbol{x})$ of *any* unbiased estimator $\hat{\boldsymbol{x}}$ by the inequality

$$\boldsymbol{E}(\boldsymbol{x}) := \mathbb{E}_{D|\boldsymbol{x}}\left[\left(\hat{\boldsymbol{x}}(D) - \boldsymbol{x}\right)\left(\hat{\boldsymbol{x}}(D) - \boldsymbol{x}\right)^{\text{T}}\right] \geqslant \boldsymbol{I}(\boldsymbol{x})^{-1}. \tag{7}$$

The Fisher information matrix for a single two- outcome measurement is generically a rank-1 matrix, and thus cannot be inverted for models with more than a single model parameter. Thus, if the Fisher information matrix is singular, as is the case here when all of the measured sequences are of the same length, the inverse is taken to be the Moore–Penrose pseudo-inverse. Since the rank of the Fisher information for multiple two-outcome measurements is limited to be at most the number of distinct measurements performed, with at least four different sequence lengths we can break the degeneracy. Since this number depends on the dimension of the model and not the underlying Hilbert space, only four measurements are required to break the degeneracy, even for systems of higher dimension than qubits.

It is often the case that we are only interested in $\tilde{p}$, the survival probability, and hence the gate fidelity, of a particular gate [13]. In this case, we can bound the error of only that parameter by looking at a single element of the error and Fisher information matrix as

$$\boldsymbol{E}(\boldsymbol{x})_{\tilde{p},\tilde{p}} \geqslant 1/\boldsymbol{I}(\boldsymbol{x})_{\tilde{p},\tilde{p}}. \tag{8}$$

To find the Fisher information for randomized benchmarking, we derive the Fisher score $\boldsymbol{q}$ of this model, conditioned on 1

$$\begin{aligned} \boldsymbol{q}(\boldsymbol{x}|1;\ m,\ \text{mode}) &= \nabla_{\boldsymbol{x}}\log\Pr(1|\boldsymbol{x};\ m,\ \text{mode}) \\ &= \Pr(1|\boldsymbol{x};\ m,\ \text{mode})^{-1}\begin{cases} \left(0,\ Amp_{\text{ref}}^{m-1},\ p_{\text{ref}}^{m},\ 1\right) & \text{reference,} \\ \left(Am\tilde{p}^{m-1}p_{\text{ref}}^{m},\ Am\tilde{p}^{m}p_{\text{ref}}^{m-1},\ p_{\text{ref}}^{m},\ 1\right) & \text{interleaved,} \end{cases} \end{aligned} \tag{9}$$

where the similar expression for the outcome '0' follows immediately. With this, we can calculate the Fisher information matrix $\boldsymbol{I}(\boldsymbol{x}) := \mathbb{E}_{D|\boldsymbol{x}}[\boldsymbol{q}(\boldsymbol{x}|D)\boldsymbol{q}(\boldsymbol{x}|D)^{\text{T}}]$, where D labels the outcomes.

Fisher information analysis is one of the most powerful tools of statistical analysis since it bounds the performance of the continuous infinity of possible estimators we could choose. However, given the difficulty of analytically computing the inverse of sums over Fisher information matrices of this form, we use numerical methods for its evaluation. In particular, QInfer [43] performs this calculation automatically, given an implementation of (9).

In experimentally relevant regimes, the task is to gain further accuracy when it is known *a priori* that the fidelity is high. To minimize the error in estimating $\tilde{p}$, we maximize the corresponding element of the Fisher information matrix. Note that, as is shown in figure 2, this optimum depends strongly on the value of A and B when $\tilde{p},\ p_{\text{ref}} \approx 1$. Critically, because randomized benchmarking requires no explicit simulation of quantum systems, it can in principle be used even in very large systems, beyond what can be studied using techniques that depend on simulation. Thus, we also consider the limit as $d \to \infty$, where for ideal measurements and unital channels, we have $B \to 0$ and $A \to 1$. In this limit, m can be explicitly optimized such that for large systems

$$\lim_{d\to\infty} m_{\text{opt}} \approx \frac{1}{1 - \tilde{F}F_{\text{ref}}}, \tag{10}$$

where m_{opt} represents the optimal sequence length. As shown in figure 2, this can grow large for $|1 - F| \to 0$, but even for fidelities near thresholds, such as $|1 - F| \approx 10^{-3}$ as considered by [9], m_{opt} remains manageable at

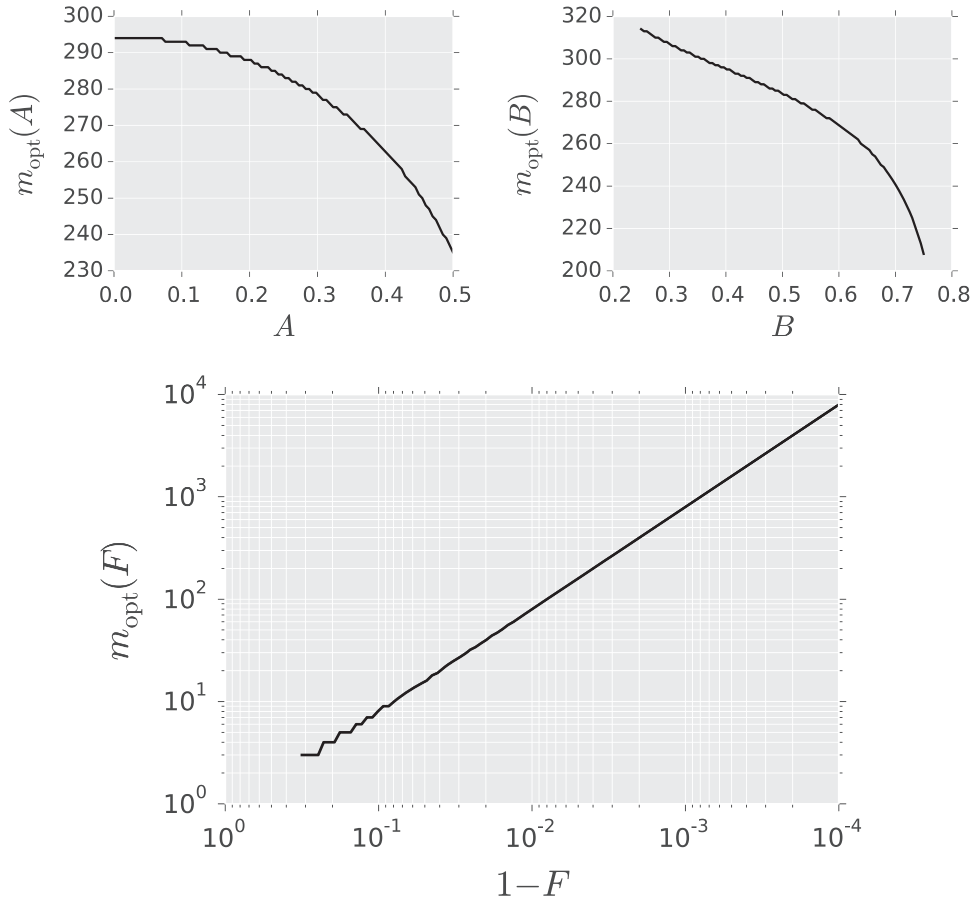

Figure 2. Optimal value of m as a function of the scale A and offset B parameters, with $\tilde{p} = 0.9988$ and $p_{\mathrm{ref}} = 0.9978$, based on the example of [9]. On the top left, $B = 0.5$. On the top right, $A = 0.25$. Below, we take the limit as $d \to \infty$ of $\lceil m_{\mathrm{opt}} \rceil$, assuming $\tilde{F} = F_{\mathrm{ref}} = F$.

about 800. This establishes that the sequence length does not grow large too quickly, providing further evidence of the utility of benchmarking even for experiments beyond the scope of tomographic methods.

The above calculation is relevant in scenarios where the parameters not of interest (that is, A and B) are known fairly well and the gate fidelity is already known to be near unity. If we have prior information that is not of this form, Bayesian analysis is better suited to the task.

The Bayesian analogue of Fisher information analysis is a straightforward generalization. We begin with a distribution $\pi(\boldsymbol{x})$, called a prior, over the parameters. Ideally, this is a faithful encoding of the the experimenter's prior information, but the following analysis works equally well for *any* distribution. In particular, given a prior distribution $\pi(\boldsymbol{x})$, the Bayesian information matrix $\boldsymbol{J}$ is then defined as [44]

$$\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}\sim\pi}[\boldsymbol{I}(\boldsymbol{x})]. \tag{11}$$

To calculate this we can perform a Monte Carlo integral over the prior by drawing samples $\boldsymbol{x} \sim \pi$ and evaluating $\boldsymbol{I}$ at each $\boldsymbol{x}$.

The Bayesian Cramer–Rao bound (BCRB) then states that the error matrix $\boldsymbol{E} := \mathbb{E}_{\boldsymbol{x},D}[(\hat{\boldsymbol{x}}(D) - \boldsymbol{x})(\hat{\boldsymbol{x}}(D) - \boldsymbol{x})^{\mathrm{T}}]$, also called the *risk*, of any estimator $\hat{\boldsymbol{x}}$ satisfies

$$\boldsymbol{E} \geqslant \boldsymbol{J}^{-1}. \tag{12}$$

The calculation of the BCRB is naturally included into the SMC algorithm, such that our approach bounds its own performance based on the best experimental data available. Moreover, contrary to the Cramer–Rao bound in equation (7), it is known that the mean of the posterior distribution minimizes the error [45]. Thus, we need not seek the optimal estimator, as it naturally arises from a representation of the posterior.

3. Numerical examples

In the numerical examples we consider here, we choose π to be a normal distribution with a mean vector $(\tilde{p}, p_{\mathrm{ref}}, A, B) = (0.95, 0.95, 0.3, 0.5)$ and equal diagonal covariances given by a deviation of $\sigma = 0.01$. The least-squares fit (LSF) estimator is seeded with an initial guess drawn from this prior, so as to fairly compare the estimators. This distribution is intersected with the hard constraints implied by definitions of the parameters, which defines the support of the prior as

$$\begin{aligned} \operatorname{supp} \pi = \{(A, B, p) : -(1 - 1/d) \leqslant A \leqslant 1,\, 0 \leqslant B \leqslant 1, \\ 0 \leqslant p \leqslant 1,\, 0 \leqslant Ap + B \leqslant 1\}. \end{aligned} \tag{13}$$

This distribution was chosen as the likelihood model is less degenerate given these constraints, such that it is easier to reason about bounds for approximately unimodal estimation strategies. Our choice of prior is not critical, however, as we will show later that our algorithm recovers well from the case in which we choose a 'bad' prior.

To demonstrate the Bayesian approach, we compare the standard LSF performance to the SMC algorithm [35], which computes estimates by updating the probability of each of a finite list of hypotheses according to Bayes' rule. In the case of randomized benchmarking, this consists of computing (6) for each hypothesis after each batch of measurements. We note that the cost of computing (6) is independent of the dimension of the system, such that randomized benchmarking explicitly avoids simulating quantum evolution with classical resources.

There are essentially two experimental design choices an experimenter can make: the length of the sequence m, and the number of repetitions K. In the first comparison, we fix the sequence length and vary K. In particular, we take all sequence lengths up to 100 for the reference signal and 50 for the interleaved signal. For each such K, we plot the mean squared error for the SMC and LSF estimators, along with the posterior variance, which provides an online estimate of the performance of SMC, and the Bayesian Cramer–Rao bound. The results, shown in figure 3, demonstrate that SMC can be used to obtain useful estimates of $\tilde{p}$ with a *few orders of magnitude* less data than is used by least-squares fitting. Moreover, this advantage becomes more pronounced as the number of shots per sequence length approaches one, such that SMC is especially useful in cases where data collection is expensive. We note that this advantage reflects both the performance of SMC itself, and the ability of SMC to take advantage of prior information: for small amounts of data, the LSF estimator chooses estimates far from the initial guess drawn from the prior distribution, while the SMC estimate instead refines the prior. Moreover, SMC can accurately characterize its own performance and can obtain significantly closer accuracy to the ultimate bound given by the BCRB. These advantages are similar to other cases in which SMC shows a large advantage over traditional fitting methods for handling data that is far from deterministic [35, 46].

In figure 3 (bottom), we show the performance of SMC and LSF when the sequence lengths m vary and the number of shots K per sequence length is fixed, demonstrating that SMC can improve upon LSF especially for very short sequences. Moreover, we see the benefit from increasing the sequence length is *minimal* compared to repeating experiments at a given sequence length near the optimum length found from the Cramer–Rao bound .

4. Benchmarking with simulated gates

Thus far in the analysis, we have used as a simulator the same zeroth-order model as is used to process and interpret the data. To demonstrate the utility of our approach in comparison with traditional LSF-based benchmarking, we now simulate gates according to a cumulant expansion, with physically realistic models. In particular, we use the superconducting model of Puzzuoli *et al* [47] together with optimal control theory [26] to generate a set of gates implementing the target unitaries $\{\mathbb{1}, X, Y, Z, H, P\}$, where H is the Hadamard gate, and where $P = |0\rangle\langle 0| + \mathrm{i}|1\rangle\langle 1|$ is the phase gate and X, Y and Z are the Pauli matrices. We then use the superoperators $\hat{\hat{S}}_U$ for implementing each target unitary U obtained from a cumulant simulation [48, 49] to sample from the likelihood function (1)[9]. An example signal is shown in figure 4.

To process these samples, we then use the zeroth-order likelihood function (6) both as a model for SMC and as a trial function for least-squares fitting. Since the actual implemented gates are known, we can compute the true parameters for comparison. In table 1, we show the true parameters, the result obtained using SMC, and the result obtained using least-squares fitting. The most important thing to note is that correct parameters are a distance 6.90 σ from the prior (meaning the true parameters are outside of the 99.999 9998% credible ellipse). This shows that even in the case when the prior information fails to accurately capture the uncertainty in the true model, SMC still can perform well, providing evidence that our accelerated methods may also be *robust*, even

[9] To ensure that the ideal action of each sequence is the identity operation, we use Gottesman–Knill simulation [53] as implemented by the QuaEC library [50] to find the inverse of the first $(m - 1)$ gates in each sequence, and then set the mth gate to be the inverse. The algorithm for implementing the simulator is described in the supplemental materials.

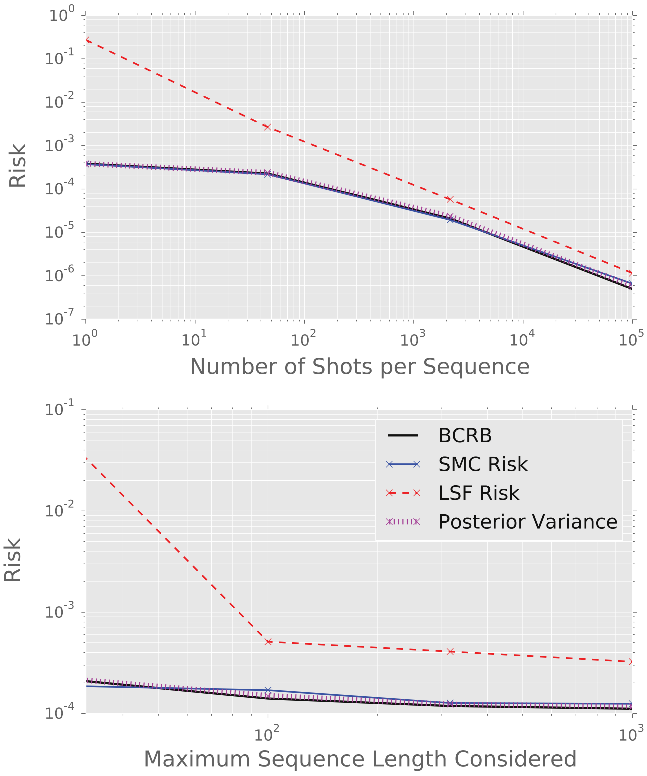

Figure 3. Comparison of the mean squared error averaged over trials (risk) achieved by sequential Monte Carlo (SMC) and least-squares fit (LSF) estimators of $\boldsymbol{x} = (\tilde{p},\ p_{\text{ref}},\ A,\ B)$, averaged over 100 trials and varied over the number of sequences per length K (top) and the number of the maximum sequences $m_{\max}$ (bottom). The Bayesian Cramer–Rao bound and posterior variance (an estimate of SMC's performance) are also shown. On top, the simulated reference signal was taken with sequence lengths $\{1,\ 2,\ldots,\ 100\}$ and the interleaved signal was taken with $m \in \{1,\ 2,\ldots,\ 50\}$. On the bottom, $K = 10^3$ samples were simulated per sequence length. For each $m_{\max}, m \in \{1,\ 11,\ 21,\ldots,\ m_{\max}\}$ was chosen.

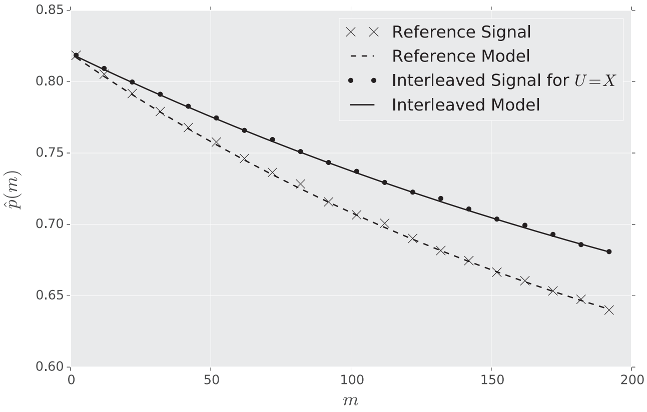

Figure 4. Example of randomized benchmarking signals collected using physical gate models. Individual points correspond to collected data with $K = 1000$ sequences per sequence length, while the curves show the true model as obtained using the zeroth-order model definitions.

when used to measure the fidelities of sets of gates with errors that are correlated between distinct gate types, or that include non-trivial unitary components[10]. We show this in more detail in figure 5, comparing the posterior and prior distributions over $\tilde{p}$ to the true and LSF-estimated values.

[10] Note that SMC did not act in a robust manner in all cases observed, but in those cases where SMC did not do well by comparison to LSF, the QInfer package was often able to warn by using the effective sample size criterion described in [35], such that the data processing could then be repeated if necessary, or such that a more appropriate prior could be chosen. This can be made more formal by appealing to model selection to decide the validity of a prior.

Table 1. Results of using SMC and least-squares fitting to estimate the fidelity of $U = X$, simulated using the superconducting qubit gate set. (Left) Bad prior from figure 5, (right) accurate prior from figure 6.

	Bad prior (40×10^3 bits)				Good prior (3×10^3 bits)			
	$\tilde{p}$	$p_{\rm ref}$	A_0	B_0	$\tilde{p}$	$p_{\rm ref}$	A_0	B_0
True	0.9983	0.9957	0.3185	0.5012	0.9983	0.9957	0.3185	0.5012
SMC estimate	0.9953	0.9971	0.2639	0.5164	0.9936	0.9976	0.3007	0.5028
LSF estimate	0.9905	0.9989	0.5525	0.2702	0.9917	0.9988	0.5266	0.2718
SMC error	0.0030	0.0014	0.0545	0.0152	0.0048	0.0019	0.0178	0.0016
LSF error	0.0078	0.0032	0.2341	0.2310	0.0066	0.0031	0.2081	0.2294
SMC relative error	0.300%	0.14%	17.12%	3.03%	0.478%	0.19%	5.58%	0.31%
LSF relative error	0.784%	0.32%	73.50%	46.08%	0.664%	0.31%	65.36%	45.78%

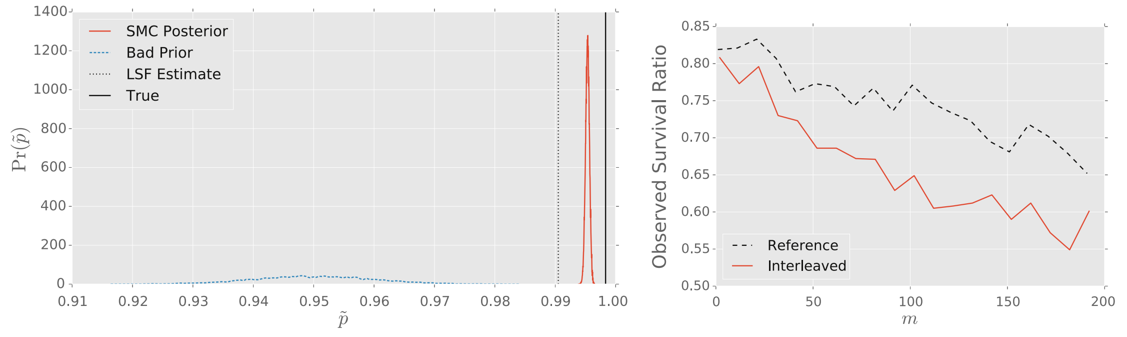

Figure 5. (Left) Comparison of prior distribution, SMC-approximated posterior, true value and LSF-estimate for $\tilde{p}$ for a single run with $K = 1000$ shots at each of $m_{\rm ref} \in \{1, 11, \ldots, 191\}$ and $m_C \in \{2, 12, \ldots, 192\}$. An intentionally inaccurate prior is used, such that the true value is approximately 6.9 standard deviations from the mean of the prior. As shown in table 1, SMC does well by comparison to LSF, even with the poorly chosen prior. (Right) Data gathered from simulation with physical-model gates.

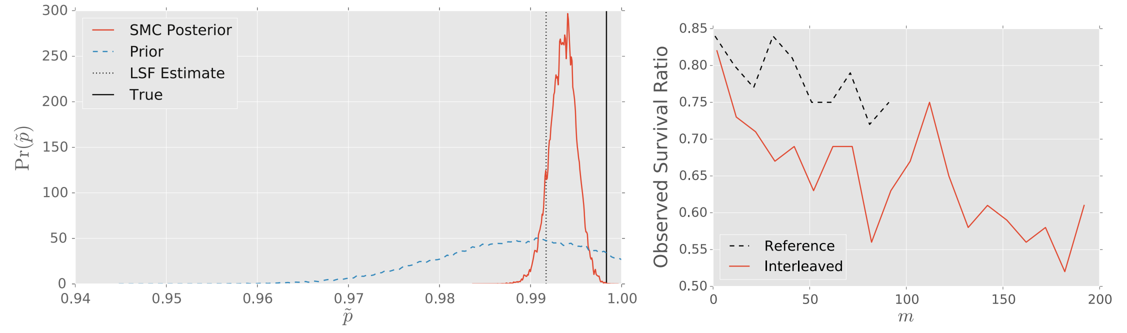

Figure 6. (Left) Comparison of prior distribution, SMC-approximated posterior, true value and LSF-estimate for $\tilde{p}$ for a single run with $K = 100$ shots at each of $m_{\rm ref} \in \{1, 11, \ldots, 91\}$ and $m_C \in \{2, 12, \ldots, 192\}$. (Right) Data gathered from simulation with physical-model gates.

Finally, in figure 6, we demonstrate the advantage of our method in the presence of physical gates together with a more reasonable prior, and using approximately ten-fold less data than in figure 5. Taken with other evidence of the robustness of SMC methods [39, 46], these results thus show that our method is useful and provides advantages in data collection costs in experimentally reasonable conditions.

We also note that LSF provides an accurate estimate of $\tilde{p}$ for the simulations with physical gates, but it appears to be at the expense of providing poor estimates for A and B. Given that the errors in $\tilde{p}$ and those in A and B are not in general uncorrelated, that LSF often provides such poor estimates of A and B makes the estimates of $\tilde{p}$ derived from LSF difficult to trust.

In this work, we have discussed the fundamental limits of the randomized benchmarking technique that are incurred due to small data sets, and have shown an algorithm that reliably saturates this optimum. In doing so, we have shown that by using SMC, with a moderate tradeoff in computational costs, one can obtain as much as two orders of magnitude improvement in estimation accuracy, such that data collection requirements are similarly reduced by as much as a 100-fold. Given the wide and expanding use of randomized benchmarking in experimental practice, this then translates to a significant performance benefit both in benchmarking, and in experimental protocols derived from benchmarking.

Acknowledgments

The numerical methods used in this paper are implemented using the QInfer, QuaEC, QuTiP 2 and SciPy libraries for Python [43, 50–52]. We thank Holger Haas for simulating the gate sets used above, and for discussions. We thank Nathan Wiebe, Ben Criger, Ian Hincks, Josh Combes, Joel Wallman, Joe Emerson and Yuval Sanders for useful discussions and insights. CG and DGC acknowledge funding from Industry Canada, CERC, NSERC and the Province of Ontario. CF acknowledges funding from NSF Grant No. PHY-1212445 and NSERC.

Appendix. Sampling variance and derivation of marginalized likelihood

In this derivation, we will focus on the zeroth-order model of Magesan *et al* [12], which gives that the average fidelity $F_g(m)$ over all sequences of length m is given by

$$F_g(m) = A_0 p^m + B_0 \tag{A.1}$$

for constants A_0 and B_0 describing the state preparation and measurement errors, and where $1 - p$ is the depolarizing strength of $W[\mathbb{E}_{C\sim\mathcal{C}_n}\hat{\hat{S}}_C]$.

We are interested in the single-shot limit, where each measurement consists of first selecting a sequence, then measuring once the survival probability for that sequence. Since this protocol makes no use of the sequence other than its length, we can describe the protocol by marginalizing over the choice of sequence, giving a probability distribution of the form $\Pr(\text{survival}|m)$, where m is a sequence length.

To derive this, we first pick a length m, and then consider the choice of sequence $\boldsymbol{i}$ out of all length-m sequences to be a random variate. Thus, there exists probabilities

$$p_{m,\boldsymbol{i}} := \Pr(\text{survival}|\boldsymbol{i}, m) = \mathrm{Tr}\left(E_\psi \hat{\hat{S}}_{\boldsymbol{i}}\left[\rho_\psi\right]\right) \tag{A.2}$$

for each individual sequence that we could have chosen, such that marginalizing over results in

$$\Pr(\text{survival}|m) = \mathbb{E}_{\boldsymbol{i}}[\Pr(\text{survival}|\boldsymbol{i}, m)]. \tag{A.3}$$

If each sequence is drawn with uniform probability, then

$$\Pr(\text{survival}|m) = \frac{1}{\left|\mathcal{C}_n\right|^m} \sum_{\boldsymbol{i}\text{ s.t. len }\boldsymbol{i}=m} p_{m,\boldsymbol{i}}. \tag{A.4}$$

We recognize this as being the average sequence fidelity $F_g(m)$ modeled by Magesan

$$\Pr(\text{survival}|m) = F_g(m) = A_0 p^m + B_0. \tag{A.5}$$

To interpret $F_g(m)$ as a likelihood directly, note that we had to consider the Bernoulli trial (single-shot) limit; had we instead taken K distinct sequences and measured each $N > 1$ times, we would have arrived at a quite different quantity

$$\hat{F}_g(m) = \sum_{k=1}^{K} \hat{F}\left(m, \boldsymbol{i}_k\right), \tag{A.6}$$

where $\hat{F}(m, \boldsymbol{i}_k)$ is the estimate of the sequence fidelity for the *particular* sequence $\boldsymbol{i}_k$.

The difference is made clear by considering an example with fixed sequence length m, and the variance for a datum $d \sim \Pr(\text{survival}|m)$ (labeling ‘survival’ as 1 and the complementary event as 0)

$$\mathbb{V}_d[d|m] = \mathbb{V}_{\boldsymbol{i}}\Big[\mathbb{E}_d[d|\boldsymbol{i}, m]] + \mathbb{E}_{\boldsymbol{i}}\Big[\mathbb{V}_d[d|\boldsymbol{i}, m\Big]\Big]. \tag{A.7}$$

The second term corresponds to the mean variance over each fixed sequence $\boldsymbol{i}_m$, and governs how well we can estimate each $F(m, \boldsymbol{i})$ individually. The first term, however, is more interesting, in that it measures the variance *over sequences* of the per-sequence survival probability $p_{m,\boldsymbol{i}} = \mathbb{E}_d[d|\boldsymbol{i}, m]$. By the argument of Wallman and Flammia [25], this is small when the fidelity being estimated is close to 1; that is, when the gates being benchmarked are very good. For gates that are farther from the ideal Clifford operators, however, or for applications such as tomography via benchmarking [23], this term is not negligible, mandating that many different sequences must be taken for $\hat{F}_g(m)$ to be a useful estimate of $F_g(m)$.

By demanding that each individual shot be drawn from an independently chosen sequence, our approach avoids this and samples from $d|m$ directly. In this way, we see a similar effect as in [32]. In particular, it is not advantageous to concentrate one's sampling on one point, but to spread samples out and gain experimental variety. Here, the one shot per sequence limit plays the role of the one sample per time-point limit in the earlier discussion.

References

[1] Wiebe N, Braun D and Lloyd S 2012 *Phys. Rev. Lett.* **109** 050505

[2] Wiebe N, Granade C, Ferrie C and Cory D 2014 *Phys. Rev.* A **89** 042314

[3] Kassal I, Whitfield J D, Perdomo-Ortiz A, Yung M-H and Aspuru-Guzik A 2011 *Annu. Rev. Phys. Chem.* **62** 185

[4] Hastings M B, Wecker D, Bauer B and Troyer M 2014 arXiv:1403.1539[quant-ph]

[5] Shor P W 1997 *SIAM J. Sci. Stat. Comput.* **26** 1484

[6] Nickerson N H, Fitzsimons J F and Benjamin S C 2014 *Phys. Rev.* X **4** 04104

[7] Borneman T W, Granade C E and Cory D G 2012 *Phys. Rev. Lett.* **108** 140502

[8] Chow J M *et al* 2014 *Nat. Commun.* **5** 4015

[9] Barends R *et al* 2014 *Nature* **508** 500

[10] Gottesman D 2010 *Proc. Symp. Appl. Math.* **68** 13–58

[11] Fowler A G, Stephens A M and Groszkowski P 2009 *Phys. Rev.* A **80** 052312

[12] Magesan E, Gambetta J M and Emerson J 2012 *Phys. Rev.* A **85** 1094–622

[13] Ryan C A, Laforest M and Laflamme R 2009 *New J. Phys.* **11** 013034

[14] Chow J M, Gambetta J M, Tornberg L, Koch J, Bishop L S, Houck A A, Johnson B R, Frunzio L, Girvin S M and Schoelkopf R J 2009 *Phys. Rev. Lett.* **102** 090502

[15] Olmschenk S, Chicireanu R, Nelson K D and Porto J V 2010 *New J. Phys.* **12** 113007

[16] Brown K R, Wilson A C, Colombe Y, Ospelkaus C, Meier A M, Knill E, Leibfried D and Wineland D J 2011 *Phys. Rev.* A **84** 030303

[17] Moussa O, da Silva M P, Ryan C A and Laflamme R 2012 *Phys. Rev. Lett.* **109** 070504

[18] Tan T R, Gaebler J P, Bowler R, Lin Y, Jost J D, Leibfried D and Wineland D J 2013 *Phys. Rev. Lett.* **110** 263002

[19] Crcoles A D, Gambetta J M, Chow J M, Smolin J A, Ware M, Strand J, Plourde B L T and Steffen M 2013 *Phys. Rev.* A **87** 030301

[20] Gambetta J M *et al* 2012 *Phys. Rev. Lett.* **109** 240504

[21] Gustavsson S, Zwier O, Bylander J, Yan F, Yoshihara F, Nakamura Y, Orlando T P and Oliver W D 2013 *Phys. Rev. Lett.* **110** 040502

[22] Kelly J *et al* 2014 *Phys. Rev. Lett.* **112** 240504

[23] Kimmel S, da Silva M P, Ryan C A, Johnson B R and Ohki T 2014 *Phys. Rev.* X **4** 011050

[24] Epstein J M, Cross A W, Magesan E and Gambetta J M 2014 *Phys. Rev.* A **89** 062321

[25] Wallman J J and Flammia S T 2014 *New J. Phys.* **16** 103032

[26] Khaneja N, Reiss T, Kehlet C, Schulte-Herbrggen T and Glaser S J 2005 *J. Magn. Reson.* **172** 296

[27] Egger D J and Wilhelm F K 2014 *Phys. Rev. Lett.* **112** 240503

[28] Schirmer S and Langbein F 2010 *2010 4th Int. Symp. on Communications Control and Signal Processing (ISCCSP)* pp 1–5

[29] Schirmer S G and Oi D K L 2009 *Phys. Rev.* A **80** 022333

[30] Huszr F and Houlsby N 2012 *Phys. Rev.* A **85** 052120

[31] Sergeevich A, Chandran A, Combes J, Bartlett S D and Wiseman H M 2011 *Phys. Rev.* A **84** 052315

[32] Ferrie C, Granade C E and Cory D G 2013 *Quantum Inf. Process.* **12** 611

[33] Shulman M D, Harvey S P, Nichol J M, Bartlett S D, Doherty A C, Umansky V and Yacoby A 2014 *Nat. Commun.* **5** 5156

[34] Combes J, Ferrie C, Cesare C, Tiersch M, Milburn G J, Briegel H J and Caves C M 2014 arXiv: 1405.5656[quant-ph]

[35] Granade C E, Ferrie C, Wiebe N and Cory D G 2012 *New J. Phys.* **14** 103013

[36] Stenberg M P V, Sanders Y R and Wilhelm F K 2014 *Phys. Rev. Lett.* **113** 210404

[37] Ferrie C 2014 *New J. Phys.* **16** 023006

[38] Wiebe N, Granade C, Ferrie C and Cory D 2014 *Phys. Rev. Lett.* **112** 190501

[39] Wiebe N, Kapoor A and Svore K 2015 *Quantum Inf. Comput.* **15** 0318

[40] Wiebe N, Granade C and Cory D G 2014 arXiv:1409.1524 [quant-ph]

[41] Magesan E *et al* 2012 *Phys. Rev. Lett.* **109** 080505

[42] Cover T M and Thomas J A 2006 *Elements of Information Theory* (Hoboken, NJ: Wiley-Interscience) ISBN 0471241954 9780471241959

[43] Granade C *et al* 2012 *QInfer: library for statistical inference in quantum information* https://github.com/csferrie/python-qinfer

[44] Dauwels J 2005 *Int. Symp. on Information Theory, 2005. ISIT 2005. Proc. (IEEE)* 425–9

[45] Lehmann E L and Casella G 1998 *Theory of Point Estimation* (Berlin: Springer) ISBN 9780387985022

[46] Ferrie C and Granade C E 2014 *Phys. Rev. Lett.* **112** 130402

[47] Puzzuoli D, Granade C, Haas H, Criger B, Magesan E and Cory D G 2014 *Phys. Rev.* A **89** 022306

[48] Kubo R 1962 *J. Phys. Soc. Japan* **17** 1100

[49] Cappellaro P, Hodges J S, Havel T F and Cory D G 2006 *J. Chem. Phys.* **125** 044514

[50] Granade C *et al* 2012 *QuaEC: quantum error correction analysis in Python* https://github.com/cgranade/python-quaec

[51] Johansson J R, Nation P D and Nori F 2013 *Comput. Phys. Commun.* **184** 1234

[52] Jones E *et al* 2001 *SciPy: open source scientific tools for Python* www.scipy.org/

[53] Aaronson S and Gottesman D 2004 *Phys. Rev.* A **70** 052328