This is the accepted manuscript made available via CHORUS. The article has been published as:

Machine Learning for Discriminating Quantum Measurement Trajectories and Improving Readout
Easwar Magesan, Jay M. Gambetta, A. D. Córcoles, and Jerry M. Chow
DOI: 10.1103/PhysRevLett.114.200501
Machine learning for discriminating quantum measurement trajectories and improving readout

Easwar Magesan, Jay M. Gambetta, A.D. Córcoles, and Jerry M. Chow
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Current methods for classifying measurement trajectories in superconducting qubit systems produce fidelities systematically lower than those predicted by experimental parameters. Here, we place current classification methods within the framework of machine learning (ML) algorithms and improve on them by investigating more sophisticated ML approaches. We find that non-linear algorithms and clustering methods produce significantly higher assignment fidelities that help close the gap to the fidelity possible under ideal noise conditions. Clustering methods group trajectories into natural subsets within the data, which allows for the diagnosis of systematic errors. We find large clusters in the data associated with $T_1$ processes and show these are the main source of discrepancy between our experimental and ideal fidelities. These error diagnosis techniques help provide a path forward to improve qubit measurements.

Maximizing the information one can extract from a physical system requires the ability to perform accurate measurements. Our goal in this paper is to provide methods for diagnosing measurement errors and increasing fidelities by using various machine learning (ML) algorithms. An important application of these techniques is in quantum information processing, where highly accurate operations and measurements are required to perform fault-tolerant information processing in the presence of noise \[1\]. We apply our methods in a superconducting qubit measurement system, and we anticipate that the generality of these techniques can be useful in a broader class of systems. Superconducting qubits (qubits) are becoming increasingly promising for experimentally demonstrating quantum protocols due to their long coherence times \[2–4\], high-fidelity multi-qubit gate operations \[5\], and the ability to perform single-shot measurements \[6–9\] in a circuit QED architecture \[10\]. In the dispersive measurement scheme of circuit quantum electrodynamics (cQED), a superconducting anharmonic oscillator, such as a transmon \[11\], is coupled to a resonator, producing a state-dependent shift of the resonator frequency. This allows for qubit measurements by driving the resonator and recording the output trajectory \[12\] in phase (I-Q) space. In practice, significant sources of random noise and systematic effects, such as $T_1$ processes (spontaneous decay), make single-shot trajectories appear complex and difficult to distinguish. There has been significant progress in reducing error-rates and measurement times \[13\] however managing, classifying, and extracting useful information from the trajectory data is extremely important for improving readout as these systems scale to larger networks.

The outline of the paper is as follows. We begin by describing the experimental system used to create the measurement data as well as how we characterize the measurement performance by the assignment fidelity $F_a$ (defined in Eq. \[1\]). We first analyze the data using the current method, which is a simplified version of linear discriminant analysis (LDA), and obtain $F_a = 0.9586$. LDA is the simplest ML classification algorithm and finds the plane that optimally separates the I-Q data trajectories under idealized assumptions of symmetric and Gaussian noise at each point in time. Next we use the quadratic extension of LDA called quadratic discriminant analysis (QDA), which allows for noise asymmetry, and find noticeable improvement in the assignment fidelity. Finally, we remove assumptions on the noise and approach the problem from a purely geometric viewpoint using support vector machines (SVM’s). The non-linear SVM provides the largest improvement giving $F_a = 0.9821$ (\textasciitilde2.4% increase), which indicates non-linear effects such as $T_1$ events are likely present. To verify this, and to understand details of the noise, we use ML clustering algorithms to find natural subclasses in the data. We find a large subclass corresponding to $T_1$ events which validates our hypothesis. Accounting for these events, we find assignment fidelities much closer to those that should be attainable in our system under the ideal noise conditions assumed for the optimality of LDA.

Let us briefly make a few points about using ML methods. First, the methods we present here can be useful in a much broader context. Any measurement scheme that produces patterns in a geometric space can potentially benefit from more advanced ML methods. Investigating the applicability to different systems will depend on the details of each situation. Second, these methods are applicable even if we are trying to improve higher fidelity measurements than those considered here. The key is that these methods can be tailored according to the types of noise present. Third, ML methods have also been applied to other problems in quantum information such as phase estimation \[14\] and asymptotic state estimation \[15\].

Our system is a single transmon qubit (Q4) coupled to a readout resonator in a lattice of four superconducting qubits \[16\] (full details of this experiment are in Ref. \[16\] and a diagram of the four-qubit setup is given in the Supplementary Material \[17\]). The main parameters of the single-qubit system are provided in Table \[1\]. The measure-
The assignment fidelity
\[ F_a = 1 - (P[0|1] + P[1|0])/2, \]
is a standard metric for characterizing how well a measurement assigns outcomes. Here \( P[0|1] \) \((P[1|0])\) is the probability of obtaining outcome “0” (“1”) given the system was prepared in \( |1\rangle \) \((|0\rangle\) and so \( F_a \in [0, 1] \). Our data consists of 51200 single-shot trajectories (shots), half initially prepared in \( |0\rangle \) and the other half in \( |1\rangle \) (denote these classes by \( C_0 \) and \( C_1 \)). The first half of the trajectories is used as a training set for classification on the
second half. The total measurement time $T$ is 2.6$\mu$s and $[0, T]$ is discretized into 163 time-points so trajectories are represented by vectors $x \in \mathbb{R}^{326}$ (let $M = 326$) where the first (last) 163 entries correspond to the real (imaginary) parts of the trajectory. Hence each $x(j)$ can be viewed as a real-valued random variable. The mean trajectories and covariance matrices for each class are denoted $\mu_0, \mu_1$ (see Fig. 3) and $\Sigma_0, \Sigma_1$ respectively.

The current method of classifying trajectories [18] is based on linear discriminant analysis (LDA) [20] and assumes the noise is highly idealized; the noise at each time is assumed to be: 1) Gaussian distributed, 2) uncorrelated with noise at other times, and 3) symmetric between $C_0, C_1$. Alternatively, this can be phrased as $\Sigma_0$ and $\Sigma_1$ are Gaussian, diagonal, and equal. Under these assumptions $x$ is associated to

$$f_{\text{LDA}}(x) = x^T [\Sigma_0^{-1} (\mu_0 - \mu_1)] ,$$

which is then assigned as 0 or 1 according to an appropriate threshold. Using this method in our experiment gives $F_a = 0.9586$.

In reality, the noise is far from these ideal conditions and one of the main goals of this paper is to deal with these more realistic scenarios. Before doing this, we can ask what fidelity we would expect from the parameters of our system if the noise did satisfy these ideal conditions. As shown in [17] this ideal fidelity, denoted $F_{\text{id}}$, is

$$F_{\text{id}} = 0.9999 \pm 0.0001 .$$

Hence there is a large discrepancy between $F_{\text{id}}$ and $F_a$ that can be due to effects such as state-preparation errors and non-Gaussian/non-linear noise. This discrepancy motivates us to investigate better methods of classifying trajectories.

Let us relax the unrealistic assumption of noise symmetry between $C_0$ and $C_1$. In this case the optimal method is quadratic discriminant analysis (QDA) [19] and each trajectory $x$ is mapped to

$$f_{\text{QDA}}(x) = -\frac{1}{2} x^T [\Sigma_0^{-1} - \Sigma_1^{-1}] x + x^T [\Sigma_0^{-1} \mu_0 - \Sigma_1^{-1} \mu_1] ,$$

and then assigned “0” or “1” according to an appropriate threshold. We computed $F_a$ using the “fitdiscr” function in Matlab for four different methods: LDAd, LDA, QDAd, and QDA (“d” represents diagonal covariance matrix and LDAd is the method of [18]). The results are in the second column of Table I. Not surprisingly, we find QDAd improves upon LDAd and allowing non-diagonal covariance matrices produces higher $F_a$. The values in Table I are the sample means from 100 repetitions. The sample variances $\sigma^2$ are $\sim 1 \times 10^{-8}$ indicating stable/reproducible results.

A value of $F_a$ for QDA was not attainable due to singular covariance matrices, which is a result of overfitting the data (having more variables than required from the correlation time in the trajectories). To remedy this, we perform dimensionality reduction using principal component analysis (PCA) [21] and find 99.9% of the variance in the data can be accounted for in a subspace of dimension $\sim 20$ (noise correlation time $\sim 85 - 180$ ns). Results with a PCA pre-processing step (using “princomp” in Matlab) are in the third column of Table I. Not surprisingly QDA provides the highest $F_a$ out of all cases considered.

<table>
<thead>
<tr>
<th>Table II. Assignment fidelities for discriminant analysis methods.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>LDAd</td>
</tr>
<tr>
<td>LDA</td>
</tr>
<tr>
<td>QDAd</td>
</tr>
<tr>
<td>QDA</td>
</tr>
</tbody>
</table>

These classification methods assume Gaussian noise and better methods are needed to deal with realistic noise. We approach this in two ways. The first is via the support vector machine (SVM) [22, 23], which requires no assumptions on the noise and can be extended to non-linear discriminating surfaces. The second is to utilize “clustering” methods in ML to naturally group the data from which we perform multi-class classification.

The linear SVM is a quadratic program based on maximizing the minimum distance of a data point to a hyperplane separating the data. The non-linear SVM is derived by defining a kernel that maps the data to a higher-dimensional space. The linear SVM in the higher-dimensional space allows for non-linear discrimination in $\mathbb{R}^M$. Due to its generality and simplicity, we chose a radial basis function kernel.

We implemented the SVM using the Matlab “fitcsvm” function and classification was repeated 100 times. The mean values with the optimal soft-margin parameter are contained in Table III (see [17] for details). Sample variances $\sigma^2$ of $F_a$ are approximately $1.9 \times 10^{-8}$. The non-linear SVM produces the highest assignment fidelity out of all methods considered, indicating non-linear effects are present. We hypothesize the main factor producing the non-linearity is $T_1$ events.

<table>
<thead>
<tr>
<th>Table III. Assignment fidelities for SVM methods.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Linear SVM</td>
</tr>
<tr>
<td>Non-linear SVM</td>
</tr>
</tbody>
</table>

Our second method for implementing a non-linear classifier combines classification and clustering algorithms. Clustering naturally groups the data into subsets and is
unsupervised" since it requires no training data. We implement k-means clustering and anticipate similar results can be obtained with other standard methods such as hierarchical clustering. We used the Matlab “kmeans” function to find \( k = 3 \) clusters in each of \( C_0 \) and \( C_1 \). We chose \( k = 3 \) to take into account both variance and systematic effects. The mean trajectories and size of the six subclasses are given in Fig. 3. \( C_0 \) is split relatively evenly into the subclasses \( S_{0,1}, S_{0,2}, S_{0,3} \) (blue colored trajectories) that mainly capture variance. We don’t see a \( C_0 \) subclass corresponding to ground state heating, however we implemented k-means for \( k = 7 \) and found a heating subclass of size \( \sim 230 \) (see Fig.2 in [17]).

\( C_1 \) has strikingly different properties as subclass \( S_{1,2} \) (green-solid line) is comprised of \( T_1 \) processes. \( S_{1,1} \) and \( S_{1,3} \) (red trajectories) are similar in size and mainly capture variance. The key point is we have found explicit shot indices for \( T_1 \) events. We verified that \( S_{1,2} \) is comprised of \( T_1 \) trajectories by performing k-means with \( k = 4 \) (see Fig.3 in [17]). From Fig. 3 we see most clearly in Q.

To perform classification, we lift the \( T_1 \) subclass \( S_{1,2} \) to a class \( C_2 \) of its own, redefine \( C_1 = S_{1,1} \cup S_{1,3}, \) keep \( C_0 \) as before, and perform multi-class classification on \( C_0, C_1, \) and \( C_2 \). We implemented four multi-class algorithms in Matlab: multi-class LDA, multi-class SVM, “TotalBoost”, and “RUSBoost”. The latter two are examples of boosting algorithms [25] and RUSBoost [26] is particularly useful since it is tailored to the case of one class (here \( C_2 \)) being significantly smaller than the rest. The results are in Table IV. We again see an increase in \( F_a \) over the discriminant analysis methods of Table II. Not surprisingly, RUSBoost provides the most significant increase. We repeated the k-means algorithm 50 times with random initializations and found it to be stable (\( \sigma^2 \) of \( F_a \sim 3 \times 10^{-6} \)).

Out of all methods considered, non-linear SVM’s produce the greatest increase in \( F_a \) (0.9586 to 0.9821). All methods are relatively stable with reproducible assignment fidelities (each method was repeated \( \sim 100 \) times; sample means of \( F_a \) are the table values and sample variances are \( \sim 1 \times 10^{-8} \)).

TABLE IV. Assignment fidelities from multi-class classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>All time-points</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-LDA</td>
<td>0.9768</td>
<td>0.9689</td>
</tr>
<tr>
<td>Multi-SVM</td>
<td>0.9784</td>
<td>0.9717</td>
</tr>
<tr>
<td>TotalBoost</td>
<td>0.9527</td>
<td>0.9413</td>
</tr>
<tr>
<td>RUSBoost</td>
<td>0.9788</td>
<td>0.9723</td>
</tr>
</tbody>
</table>

While we have improved \( F_a \) to 0.9821, we are still far from \( F_{id} = 0.9999 \). We hypothesize much of the remaining discrepancy comes from \( T_1 \) events. To investigate this we propose the simple diagnostic test of replacing each \( T_1 \) event from the k-means algorithm with a random trajectory from \( S_{1,2} \cup S_{1,3} \). This provides a measure of \( F_a \) when \( T_1 \) is negligible. The means of 100 samples for each classification method are in Table IV (variances are \( \sim 1 \times 10^{-8} \)). Non-linear SVM produces the highest value of \( F_a \) however for all methods \( F_a > 0.99 \), which
is more consistent with $F_{id} = 0.9999$. This confirms $T_1$ events are the significant reason for not reaching $F_{id}$.

TABLE V. Assignment fidelities with replacement of $T_1$ events.

<table>
<thead>
<tr>
<th>Method</th>
<th>All time-points</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.9920</td>
<td>0.9909</td>
</tr>
<tr>
<td>LDA</td>
<td>0.9921</td>
<td>0.9928</td>
</tr>
<tr>
<td>QDA</td>
<td>0.9918</td>
<td>0.9908</td>
</tr>
<tr>
<td>QDA</td>
<td></td>
<td>0.9927</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>0.9936</td>
<td>0.9943</td>
</tr>
<tr>
<td>Non-linear SVM</td>
<td>0.9945</td>
<td>0.9949</td>
</tr>
</tbody>
</table>

One attempt to reduce the significance of $T_1$ is to reduce $T$, however this implies the trajectories will spend less time near their steady states and noise variance will dominate. To observe this, we truncated the trajectories to different $T$ and calculated $F_n$ using the non-linear SVM. From Fig. [6] $T = 2.6\mu s$ appears close to optimal. A much shorter measurement time of $\sim 1.2\mu s$ is needed to achieve an assignment fidelity of 0.9586, which we recall is the value obtained using the simplest ML method of [18]. This is a strong message that better classifiers can allow for shorter measurement times. Longer measurement times than the current $2.6\mu s$ decrease $F_n$ due to an increase in $T_1$ events.

![FIG. 6. Varying measurement time (color online).](image)

To conclude, we have utilized ML to understand and improve the readout in a superconducting system. More sophisticated classification algorithms can potentially allow for shorter measurement times and increase assignment fidelities. Non-linear SVM’s provided the largest increase in assignment fidelity, 0.9586 to 0.9821 ($\sim 2.4\%$). Clustering helped diagnose the prevalence of systematic effects by finding clusters in the data corresponding to single-shot identification of heating and $T_1$ effects. We verified $T_1$ events are a significant source of error as the assignment fidelity increases to 0.9945 when the $T_1$ cluster is replaced with typical trajectories. This is more consistent with the ideal fidelity and the remaining discrepancy can be due to effects such as heating and state-preparation errors. Moving forward, we expect these methods will help provide insight for improving readout, especially when non-linear and non-Gaussian effects are present.

We acknowledge support from ARO under contract W911NF-14-1-0124 and IARPA under contract W911NF-10-1-0324. We acknowledge helpful discussions with Oliver Dial, Stefan Filipp, Blake Johnson, Jim Rozen, Colm Ryan, Marcus Silva, and Matthias Steffen.