Practical Trainable Temporal Postprocessor for Multistate Quantum Measurement

Saeed A. Khan,[1,*] Ryan Kaufman,[2] Boris Mesits,[2] Michael Hatridge,[2] and Hakan E. Türeci[1]

[1] *Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA*

[2] *Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania, USA*

(Received 15 November 2023; revised 25 March 2024; accepted 20 May 2024; published 21 June 2024)

We develop and demonstrate a trainable temporal postprocessor (TPP) harnessing a simple but versatile machine learning algorithm to provide optimal processing of quantum measurement data subject to arbitrary noise processes for the readout of an arbitrary number of quantum states. We demonstrate the TPP on the essential task of qubit state readout, which has historically relied on temporal processing via matched filters in spite of their applicability for only specific noise conditions. Our results show that the TPP can reliably outperform standard filtering approaches under complex readout conditions, such as high-power readout. Using simulations of quantum measurement noise sources, we show that this advantage relies on the TPP's ability to learn optimal linear filters that account for general quantum noise correlations in data, such as those due to quantum jumps, or correlated noise added by a phase-preserving quantum amplifier. Furthermore, we derive an exact analytic form for the optimal TPP weights: this positions the TPP as a linearly scaling generalization of matched filtering, valid for an arbitrary number of states under the most general readout noise conditions, all while preserving a training complexity that is essentially negligible in comparison with that of training neural networks for processing temporal quantum measurement data. The TPP can be autonomously and reliably trained on measurement data and requires only linear operations, making it ideal for field-programmable gate array implementations in circuit QED for real-time processing of measurement data from general quantum systems.

DOI: 10.1103/PRXQuantum.5.020364

I. INTRODUCTION

High-fidelity quantum measurement is essential for any quantum information processing scheme, from quantum computation to quantum machine learning. However, while measurement optimization has focused on quantum hardware advancements [1–3], several modern experiments operate in regimes where optimal hardware conditions are difficult to sustain or—for machine learning with general quantum systems [4–8]—may not always be known. For example, in the push towards higher qubit readout fidelities with complex multiqubit processors in circuit QED (cQED), optimization of individual readout resonators becomes increasingly difficult. More importantly, finite qubit coherence means that simply extending the measurement duration is not a viable option to increase fidelity: faster and, hence, higher-power measurements are needed. However, these readout powers are associated with enhanced qubit transitions, leading to the T_1 versus $\bar{n}$ problem [9–15] and excitation to higher states [14,16,17] outside the computational subspace. Machine learning with quantum devices operating in unconventional regimes allows for an even broader range of complex dynamics. Quantum measurement data obtained under these conditions cannot be expected to be optimally analyzed with use of schemes built for more standard readout paradigms [18]. Therefore, a practical approach to extract the maximum information possible from such data is timely.

In this paper, we demonstrate a machine-learning scheme to optimally process quantum measurement data for completely general quantum state classification tasks. For the most common such task of single-shot qubit state readout, standard postprocessing of measurement records has remained relatively unchanged (with some exceptions [19,20]): data are filtered with use of a "matched filter" (MF) constructed from the sample mean of measurement records for two states to be distinguished (for example, state $|e\rangle$ or state $|g\rangle$ of a qubit). Crucially, the MF thus defined applies only to binary classification, and much more restrictively is optimal only for idealized conditions

*Corresponding author: saeedk@princeton.edu

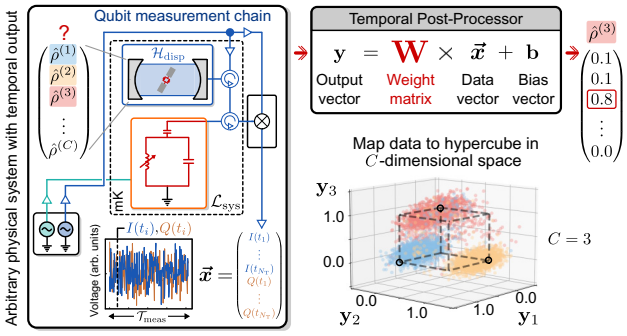


FIG. 1. TPP for multistate classification using quantum measurement data, demonstrated for dispersive qubit readout in cQED. The objective is to process temporal data corresponding to an unknown state (indexed σ) of an arbitrary physical system—here the state of a qubit in a quantum measurement chain—to estimate the true label σ with maximum accuracy. The TPP approach uses a set of weights $\mathbf{W}$ and biases $\mathbf{b}$ to map the vector $\vec{\boldsymbol{x}}$ of measured data, comprising an instance of N_{O} observables each a time series of length N_{T}, to the corners of a hypercube in C-dimensional space. Optimal values of $\mathbf{W}$ and $\mathbf{b}$ are learned by training to realize this mapping with minimal error, in a least-squares sense. Scatter plots shown in $C = 3$ dimensional space are data from real qubit $p \in \{e, g, f\}$ readout after applying the TPP.

under which the readout signal is subject to Gaussian white (i.e., uncorrelated) noise processes [21]. In many deployments where complex conditions prevail (such as multiqubit readout) an even simpler and less optimal boxcar filter is used, due to the ease of its construction. Our approach harnesses machine learning to provide a model-free trainable temporal postprocessor (TPP) of quantum measurement data under the most general noise conditions, and for an arbitrary number of states of a generic measured quantum system (see Ref. [22] for the source code). We test our approach by applying it to the experimental readout of distinct qubits across a range of measurement powers. Our results demonstrate that the TPP reliably outperforms the standard MF whenever measured data exhibit nontrivial temporal correlations, including those of a quantum origin. We find an important such regime that has attracted significant attention recently to be that of high-power readout [16,23–26]; here we experimentally show that the TPP can provide a reduction in errors by up to 30% in certain cases. Furthermore, the TPP achieves this improvement while requiring only linear weights applied to quantum measurement data (see Fig. 1): this makes it compatible with field-programmable gate array (FPGA) implementations for real-time hardware processing, and exacts a lower training cost [27,28] than neural network–based machine-learning schemes [20,29,30].

Machine learning has already been established as a powerful approach to *classical* temporal data processing, providing state-of-the-art fidelity in tasks such as time series prediction [31], and forecasting [32–34] and control [35] of chaotic systems. Adapting this approach to quantum state classification as we do here requires its application to time-evolving *quantum* signals. Signals extracted from the readout of quantum systems are often dominated by noise, making their processing distinct from that required of typical data from classical systems. More importantly, the noise in such signals can arise from truly quantum-mechanical sources, such as stochastic transitions between states of a multilevel atom (quantum "jumps") or vacuum fluctuations in quantum modes. A key finding of our work is that the TPP is able to learn from precisely these quantum noise correlations in data extracted from quantum systems to increase classification fidelity. To uncover this essential principle of TPP learning, we first develop an interpretation of the TPP as the application of optimal filters to quantum measurement data. This provides a framework to quantify and visualize what is "learned" by the TPP from a given dataset. Secondly, the TPP is tested on simulated quantum measurement datasets with use of stochastic master equations, where quantum noise sources and hence their correlation signatures in measured data can be precisely controlled.

Using simulated datasets where all noise sources contribute additive Gaussian white noise—a reasonable assumption for measurement chains under asymptotically ideal conditions—we show that the TPP provides filters that reduce *exactly* to the matched filter for binary classification. More importantly, as the TPP is valid for the classification of any number of states, it provides the generalization of matched filters for arbitrary state classification. We then provide a systematic analysis of the TPP applied to quantum measurement with more complex quantum noise sources, such as quantum amplifiers adding correlated quantum noise, or noise due to state transitions. In such scenarios the TPP provides filters adapted to the noise characteristics: we also provide an efficient semianalytic form for these general TPP filters, which can deviate substantially from filters learned under the white noise assumption and crucially outperform the latter in qubit classification. By learning from quantum noise correlations, the TPP therefore utilizes a characteristic of quantum measurement data inaccessible to postprocessing schemes relying on noise-agnostic matched filtering methods.

The established learning principles provide a structure and interpretability to the general applicability of the TPP which enhances its practical utility. First, the exact mapping to matched filters under appropriate noise conditions places the TPP on firm footing, guaranteed to perform at least as well as these baseline methods. Secondly, and much more importantly, the TPP's ability to learn from noise (crucially, quantum noise) renders it able to then beat the MF when noise conditions change. This theoretical adaptability becomes practical due to the TPP's straightforward training procedure, which is also ideal for

autonomous repeated calibrations, which are necessary on even industrial-grade quantum processors [36–38]. Ultimately, the trainable TPP could provide an ideal component to optimally process quantum measurement data from general quantum devices used for machine learning, which could exhibit exotic quantum noise characteristics.

The rest of this paper is organized as follows. In Sec. II we introduce the TPP framework to multistate classification: a model-free supervised-machine-learning approach that can be applied to the classification of arbitrary time series. We also introduce the task used to demonstrate the TPP—dispersive qubit readout in the cQED architecture—and standard approaches currently used for this task. In Sec. III we draw connections between the TPP approach and these standard filtering-based approaches to qubit state measurement, and provide the TPP's generalization of matched filtering to arbitrary states. In Sec. IV we apply the developed TPP framework to experimental data for qubit readout, showing that it can outperform standard matched filtering at the high measurement powers relevant for high-fidelity readout. Section V explores the learning principles that enable the TPP to be more effective than standard matched filters using controlled simulations. We conclude with a discussion of the general applicability of the TPP for quantum state classification and temporal processing of quantum measurement data.

II. TRAINABLE TEMPORAL POSTPROCESSOR FOR MULTISTATE CLASSIFICATION

To provide an overview of its key features, we first introduce the mathematical framework underpinning our trainable TPP, which is defined as follows. We consider N_{O} continuously measured observables, each measurement yielding a time series of length N_{T}. All measured data corresponding to an unknown state with index σ can be compiled into the vector $\vec{\boldsymbol{x}}^{(\sigma)}$, which thus exists in the space $\vec{\boldsymbol{x}}^{(\sigma)} \in \mathbb{R}^{N_{\mathrm{O}}N_{\mathrm{T}}}$, where $\vec{(\cdot)}$ specifies vectors containing vectorized *temporal* data; examples are provided shortly (see also Fig. 1).

Formally, operation of the TPP is then described as an input-output transformation, mapping a vector $\vec{\boldsymbol{x}}^{(\sigma)}$ from the space of measured data, $\mathbb{R}^{N_{\mathrm{O}}N_{\mathrm{T}}}$, to a vector $\mathbf{y} \in \mathbb{R}^{C}$ in the space of class labels; the scalar predicted class label σ^{est} is given by an operation $F[\cdot]$ on this vector $\mathbf{y}$, so the complete transformation is

$$\sigma^{\mathrm{est}} = F[\mathbf{y}] = F\left[\mathbf{W}\vec{\boldsymbol{x}}^{(\sigma)} + \mathbf{b}\right]. \tag{1}$$

Crucially, the TPP transformation—defined by a trainable matrix of weights $\mathbf{W} \in \mathbb{R}^{C\times N_{\mathrm{O}}N_{\mathrm{T}}}$ and a trainable vector of biases $\mathbf{b} \in \mathbb{R}^{C}$—is *linear*. Machine learning using only linear trainable weights has shown remarkable success in time-dependent supervised-machine-learning tasks to map time series to a dynamically evolving target function, although with a focus on classical data with weak noise [27,28]. Here we adapt this framework to processing of temporal measurement data from a quantum system and with a *time-independent* target, as is relevant for initial state classification [21].

TABLE I. Summary of components of the TPP learning framework and their dimensions.

Component	Symbol	Dimensions
TPP output	$\mathbf{y}$	$\mathbb{R}^{C}$
Weights	$\mathbf{W}$	$\mathbb{R}^{C\times(N_{\mathrm{O}}N_{\mathrm{T}})}$
Data	$\vec{\boldsymbol{x}}$	$\mathbb{R}^{N_{\mathrm{O}}N_{\mathrm{T}}}$
Bias	$\mathbf{b}$	$\mathbb{R}^{C}$
Data means, state p	$\vec{\boldsymbol{s}}^{(p)}$	$\mathbb{R}^{N_{\mathrm{O}}N_{\mathrm{T}}}$
Noise process, state p	$\vec{\boldsymbol{\zeta}}^{(p)}$	$\mathbb{R}^{N_{\mathrm{O}}N_{\mathrm{T}}}$
"Gram" matrix	$\mathbf{G}$	$\mathbb{R}^{(N_{\mathrm{O}}N_{\mathrm{T}})\times(N_{\mathrm{O}}N_{\mathrm{T}})}$
Correlation matrix	$\mathbf{V}$	$\mathbb{R}^{(N_{\mathrm{O}}N_{\mathrm{T}})\times(N_{\mathrm{O}}N_{\mathrm{T}})}$

More precisely, $\mathbf{W}$ and $\mathbf{b}$ are both learned from sampled data $\vec{\boldsymbol{x}}^{(p)}$ with *known* labels p (C in total) in a supervised learning framework. The target $\mathbf{y} \in \mathbb{R}^{C}$ for any instance of $\vec{\boldsymbol{x}}^{(p)}$ is taken to be a vector with only one nonzero element—a single 1 at index p, defining a corner of a C-dimensional hypercube (referred to as one-hot encoding, see Fig. 1). Then the optimal $\mathbf{W}^{\mathrm{opt}}$ and $\mathbf{b}^{\mathrm{opt}}$ minimize a least-squares cost function to achieve this target with minimal error:

$$\{\mathbf{W}^{\mathrm{opt}}, \mathbf{b}^{\mathrm{opt}}\} = \underset{\mathbf{W},\mathbf{b}}{\mathrm{argmin}} ||\mathbf{Y} - (\mathbf{W}\mathbf{X} + \mathbf{b})||^{2}. \tag{2}$$

Here $\mathbf{X}$ is the matrix containing the complete training dataset, comprising N_{train} instances of $\vec{\boldsymbol{x}}^{(p)}$ for each class p, while $\mathbf{Y}$ is the corresponding set of targets (see Appendix C for full training details).

A distinguishing feature of the TPP framework among other ML paradigms is that its optimization is convex and hence guaranteed to converge. The function $F[\cdot]$ used to map the TPP output to an estimated class label is *untrained*, and hence does not effect the training complexity; it is often taken to be the argmax$\{\cdot\}$ function that extracts the position of the largest element in $\mathbf{y}$. However, it can also be a more general classifier, such as a Gaussian discriminator (clarified shortly). The dimensions of the various components making up the TPP framework are summarized in Table I.

A. Learning from noise correlations

While Eq. (1) presents a formal mathematical formulation of the TPP framework in the machine-learning context, we can develop further understanding of how the TPP learns from data to enable classification. To this end, we first note that these stochastic measurement data can be

written in the very general form

$$\vec{x}^{(\sigma)} = \vec{s}^{(\sigma)} + \vec{\xi}^{(\sigma)}. \tag{3}$$

Here $\vec{\xi}^{(\sigma)}$ describes the stochasticity of the measured data: most importantly, we are interested in data where $\vec{\xi}^{(\sigma)}$ will be dominated by contributions from quantum noise sources. We take the noise process to have zero mean, $\mathbb{E}[\vec{\xi}_j^{(\sigma)}] = 0$, where $\mathbb{E}[\cdot]$ describes ensemble averages over distinct noise realizations (obtained for distinct measurements). Then $\vec{s}^{(\sigma)} = \mathbb{E}[\vec{x}^{(\sigma)}]$ is simply the sample mean of the measured data traces for state σ. Crucially, the noise is characterized by nontrivial second-order temporal correlations, which we define as $\mathbf{\Sigma}_{jk}^{(\sigma)} = \mathbb{E}[\vec{\xi}_j^{(\sigma)}\vec{\xi}_k^{(\sigma)}]$. Higher-order correlations of the noise can also be generally nonzero, but they are not explicitly analyzed here due to the TPP's use of a quadratic loss function.

The use of a least-squares cost function in Eq. (2) is now crucial: it means that a closed form of the optimal weights $\mathbf{W}^{\text{opt}}$ and biases $\mathbf{b}^{\text{opt}}$ learned by the TPP can be obtained (see Appendix D). Furthermore, the form of Eq. (3) allows us to write these learned weights and biases as

$$\begin{pmatrix}\mathbf{W}^{\text{opt}} & \mathbf{b}^{\text{opt}}\end{pmatrix} = \mathbf{M}\mathbf{D}^{-1}. \tag{4}$$

Here $\mathbf{M}$ is a matrix that depends only on the mean traces (full form in Appendix D). In contrast, $\mathbf{D}$ is the matrix of second-order moments,

$$\mathbf{D} = \begin{pmatrix} \mathbf{G}+\mathbf{V} & \sum_c \vec{s}^{(c)} \\ \sum_c (\vec{s}^{(c)})^T & C \end{pmatrix}, \tag{5}$$

which depends on the "Gram" matrix of mean traces, $\mathbf{G} = \sum_c \vec{s}^{(c)}(\vec{s}^{(c)})^T$, but also on the temporal correlations via the matrix $\mathbf{V} \equiv \sum_c \mathbf{\Sigma}^{(c)}$. Both these quantities emerge naturally in the analysis of the resolvable expressive capacity of physical systems that are subject to noise [39]. Here, Eq. (4) implies that weights learned by the TPP are not determined only by data *means* via $\mathbf{G}$ but are also sensitive to temporal *correlations* through $\mathbf{V}$. This simple feature will distinguish the TPP from standard classification approaches, a result we demonstrate in the rest of our analysis.

B. Quantum noise in dispersive qubit readout

We demonstrate the utility of the TPP framework for contemporary cQED applications by focusing on readout of dispersive qubit-cavity systems. However, we emphasize that the TPP is model-free: it can process data $\vec{x}$ generated by an arbitrary physical system, without any knowledge of its underlying physical model. Nevertheless, we introduce a simplified theoretical model of dispersive qubit-cavity systems below to erect a foundation for the interpretability of the TPP [40]. First, this enables us to identify the sources of quantum noise at play in dispersive qubit readout. More importantly, we use this model to generate benchmarking datasets with controlled, practically relevant quantum noise characteristics: the TPP's application to these datasets with known temporal correlations in Secs. III and V allows us to interpret its learning principles. The ultimate test for the TPP is still in its application to real qubit readout data, in Sec. IV.

The standard quantum measurement chain for heterodyne readout of a multilevel artificial atom (here, a transmon) dispersively coupled to a readout cavity is depicted schematically in Fig. 1 and can be modeled via the stochastic master equation (SME)

$$d\hat{\rho}_c = \mathcal{L}_{\text{sys}}\hat{\rho}_c\, dt + \mathcal{L}_{\text{envt}}\hat{\rho}_c\, dt + \mathcal{L}_{\text{meas}}[dW]\hat{\rho}_c. \tag{6}$$

Here the Liouvillian superoperator $\mathcal{L}_{\text{sys}}$ defines the quantum system whose states are to be read out. For dispersive qubit readout, $\mathcal{L}_{\text{sys}}\hat{\rho} = -i[\hat{\mathcal{H}}_{\text{disp}}, \hat{\rho}]$, where the dispersive Hamiltonian $\hat{\mathcal{H}}_{\text{disp}}$ for a multilevel transmon takes the form (for cavity operators in the interaction frame with respect to an incident readout tone at frequency ω_d and with our setting $\hbar = 1$)

$$\hat{\mathcal{H}}_{\text{disp}} \simeq \sum_p \omega_p |p\rangle\langle p| - \Delta_{da}\hat{a}^\dagger\hat{a} + \sum_p \chi_p \hat{a}^\dagger\hat{a}|p\rangle\langle p|. \tag{7}$$

Here $\Delta_{da} = \omega_d - \omega_a$ is the detuning between the cavity and the readout tone, while χ_p is the dispersive shift per photon when the artificial atom is in state $|p\rangle$ [41,42]. Unfortunately, the artificial atom can undergo transitions from its initial state to unmonitored loss channels, which can reduce readout fidelity; all losses through such channels are described by the general Liouvillian $\mathcal{L}_{\text{envt}}$.

The final superoperator $\mathcal{L}_{\text{meas}}$ defines measurement chain components that are actively monitored to read out the state of the quantum system of interest. Here we consider continuous heterodyne monitoring of a single quantum mode of the measurement chain, generally labeled $\hat{d}$. In the simplest case, $\mathcal{L}_{\text{meas}}$ defines readout of the cavity itself (then, $\hat{d} \to \hat{a}$); however, it can also describe the dynamics (coherent or otherwise) of any other monitored quantum devices in the measurement chain. The most pertinent example is readout of the signal mode of an (ideally linear) quantum-limited amplifier that follows the dispersive qubit-cavity system via an intermediate circulator, as shown schematically in Fig. 1. Most generally, $\mathcal{L}_{\text{meas}}$ can describe the monitoring of several modes of a general quantum nonlinear processor that is embedded in the measurement chain [5]. Crucially, $\mathcal{L}_{\text{meas}}$ must include a stochastic component (indicated by the Wiener increment dW), describing measurement-conditioned dynamics

of the dispersive qubit-cavity system under such continuous monitoring (see Appendix B).

For a qubit in the (*a priori* unknown) initial state $|\sigma\rangle$ before measurement, continuous monitoring of the measurement chain then yields a single "shot" of heterodyne records $\{I^{(\sigma)}(t), Q^{(\sigma)}(t)\}$ contingent on this state σ. The complexity of this readout task can be appreciated given the form of raw heterodyne records even under a simplified theoretical model:

$$I^{(\sigma)}(t_i) = \sqrt{\kappa}\langle \hat{X}^{(\sigma)}(t_i)\rangle + \xi_I(t_i) + \xi_I^{\mathrm{QM}}(t_i) + \xi_I^{\mathrm{cl}}(t_i), \quad (8a)$$

$$Q^{(\sigma)}(t_i) = \sqrt{\kappa}\langle \hat{P}^{(\sigma)}(t_i)\rangle + \xi_Q(t_i) + \xi_Q^{\mathrm{QM}}(t_i) + \xi_Q^{\mathrm{cl}}(t_i). \quad (8b)$$

We consider discretized temporal indices t_i, for $i \in [N_{\mathrm{T}}]$ and $N_{\mathrm{T}} = \mathcal{T}_{\mathrm{meas}}/\Delta t$, where $\mathcal{T}_{\mathrm{meas}}$ is the total measurement time and Δt is the sampling time set by the digitizer. Heterodyne measurement is intended to probe the expectation values $\langle \hat{X}^{(\sigma)}(t_i)\rangle, \langle \hat{P}^{(\sigma)}(t_i)\rangle$ of canonical quadratures $\hat{X} = \frac{1}{\sqrt{2}}(\hat{d} + \hat{d}^\dagger)$, $\hat{P} = -\frac{i}{\sqrt{2}}(\hat{d} - \hat{d}^\dagger)$ of the monitored mode $\hat{d}$; however, any individual measurement record is obscured by noise ξ from various sources.

Vacuum noise $\xi_I(t_i), \xi_Q(t_i)$ is associated with heterodyne measurement of even an empty cavity, and is modeled as zero-mean Gaussian white noise,

$$\mathbb{E}[\xi_{I,Q}(t_i)] = 0, \quad \mathbb{E}[\xi_{I,Q}(t_i)\xi_{I,Q}(t_j)] = \frac{1}{\Delta t}\delta_{ij}\,\delta_{I,Q}. \quad (9)$$

More importantly, $\xi_I^{\mathrm{QM}}(t_i), \xi_Q^{\mathrm{QM}}(t_i)$ describe quantum noise contributions to measurement records, whose origin is intrinsically tied to the nature of quantum measurement. The measurement of a quantum system imposes an evolution of its state, so a given measurement affects the outcome of subsequent measurements. This effect is described via a measurement-conditioned stochastic quantum state $\hat{\rho}_c$ (referred to as a *quantum trajectory*), which is distinct from the unconditional quantum state $\hat{\rho}$ formally obtained from ensemble-averaging over repeated measurements. Consequently, for any given measurement instance, observables such as the conditional quadrature expectation $\langle \hat{X}^{(\sigma)}(t_i)\rangle_c = \mathrm{Tr}\{\hat{X}\,\hat{\rho}_c^{(\sigma)}(t_i)\}$ under heterodyne monitoring can deviate from the unconditional ensemble average $\langle \hat{X}^{(\sigma)}(t_i)\rangle$; this difference, given by $\xi_I^{\mathrm{QM}}(t_i) = \langle \hat{X}^{(\sigma)}(t_i)\rangle_c - \langle \hat{X}^{(\sigma)}(t_i)\rangle$, manifests itself as quantum noise. These terms include amplified quantum fluctuations when one is measuring the output field from a quantum amplifier (see Sec. V A) or the influence of quantum jumps in the measured cavity field due to transitions of the dispersively coupled qubit (see Sec. V B). Finally, $\xi_I^{\mathrm{cl}}(t_i), \xi_Q^{\mathrm{cl}}(t_i)$ describe classical noise contributions to measurement records, for example, noise added by classical HEMT amplifiers. While the statistics of this noise may take different forms, it is formally distinct from heterodyne measurement noise, as it has no associated stochastic measurement superoperator in Eq. (6).

The objective of the qubit readout task is then to use noisy *single-shot* [43] temporal measurement data to obtain an estimated class label σ^{est} that is ideally equal to the true class label σ. Within the TPP framework, $N_{\mathrm{O}} = 2$ and $\vec{\boldsymbol{x}}^{(\sigma)} = \begin{pmatrix} \vec{I}^{(\sigma)} \\ \vec{Q}^{(\sigma)} \end{pmatrix}$, where $\vec{I}_i = I(t_i)$. The noise $\vec{\boldsymbol{\zeta}}$ in Eq. (3) then contains the terms ξ, ξ^{QM}, and ξ^{cl}. However, before describing TPP results, we first briefly review standard approaches to qubit state classification.

C. Standard postprocessing for binary qubit state readout: Matched filters

The standard classification paradigm in cQED to obtain σ^{est} from raw heterodyne records would formally be described as a filtered Gaussian discriminant analysis (FGDA) in contemporary learning theory [44], sometimes also referred to as a "Gaussian mixture model". This comprises two stages: (1) temporal filtering of each measured quadrature and (2) assignment of a class label to filtered quadratures that maximizes the likelihood of their observation among all C classes as determined by a Gaussian probability density function. Formally, this procedure can be written as

$$\sigma^{\mathrm{est}} = G\left[\sum_i \begin{pmatrix} h_I(t_i) I^{(\sigma)}(t_i) \\ h_Q(t_i) Q^{(\sigma)}(t_i) \end{pmatrix}\right] = G\left[\begin{pmatrix} \vec{h}_I^T \vec{I}^{(\sigma)} \\ \vec{h}_Q^T \vec{Q}^{(\sigma)} \end{pmatrix}\right]. \quad (10)$$

The function $G[\cdot]$ then assigns class labels according to the aforementioned Gaussian discriminator.

A fact seldom mentioned explicitly is that both the temporal filters and the Gaussian discriminator must be constructed with use of a calibration dataset, analogous to the training phase of the TPP: a set of N_{train} heterodyne records obtained when the initial qubit states are known under controlled initialization protocols. For example, for the most commonly considered case of binary qubit state classification to distinguish states $|e\rangle$ and $|g\rangle$, and under the assumption that the noise in heterodyne records is additive Gaussian white noise, an optimal filter is known: the matched filter [21,45,46]. The empirical matched filter is constructed from the calibration dataset, where (n) indexes distinct records, via

$$\vec{h}_I = \frac{1}{N_{\mathrm{train}}} \sum_{n=1}^{N_{\mathrm{train}}} \left(\vec{I}_{(n)}^{(e)} - \vec{I}_{(n)}^{(g)}\right), \quad (11)$$

with $\vec{h}_Q$ defined analogously for $I \to Q$. The function $G[\cdot]$ requires the fitting of Gaussian profiles to measured probability distributions of known classes, and hence uses means and variances estimated from calibration data.

While a Gaussian discriminant analysis can be applied to classification of an arbitrary number of states C and beyond white noise constraints, the choice of an optimal temporal filter in these more general situations is not straightforward [47]. Because of its ease of construction, a matched filter akin to Eq. (11), or an even more rudimentary boxcar filter (a uniform filter that is nonzero only when the measurement signal is ON) is often deployed, regardless of the complexity of the noise conditions (for example, when qubit decay is significant and more optimal filters can be found [21]). We will show how the TPP approach provides a natural generalization of matched filtering to multistate classification and furnishes a trainable classifier that can generalize to more complex noise environments.

III. TPP LEARNING AS OPTIMAL FILTERING: GENERALIZED MATCHED FILTERS

To understand how the TPP generalizes standard matched filtering approaches, we first show an important connection between the two schemes. Note that the learned matrix of weights $\mathbf{W}^{\text{opt}} \in \mathbb{R}^{C\times N_O N_T}$ can be equivalently expressed as

$$\mathbf{W}^{\text{opt}} = \begin{pmatrix} \vec{f}_1^{\,T} \\ \vdots \\ \vec{f}_C^{\,T} \end{pmatrix}, \tag{12}$$

where $\vec{f}_k \in \mathbb{R}^{N_O N_T}$ for $k \in [C]$. With this parameterization, Eq. (1) for the k th component of the vector $\mathbf{y}$ can be rewritten as

$$\mathbf{y}_k = \vec{f}_k^{\,T}\vec{x} + \mathbf{b}_k, \quad k \in [C]. \tag{13}$$

When Eq. (13) is compared against Eq. (10), the interpretation of $\vec{f}_k$ becomes clear: this set of weights can be viewed as a temporal filter applied to the data $\vec{x}$. TPP-based classification can therefore be interpreted as the application of C filters (one for each k) to obtain the estimated label σ^{est}. The optimal $\mathbf{W}^{\text{opt}}$ therefore defines the optimal filters that enable this estimation with minimal error. The use of C optimal filters for a C-state classification task indicates the linear scaling of the TPP approach with the complexity of the task.

Remarkably, the optimal $\mathbf{W}^{\text{opt}}$ given by Eq. (4), and hence the C optimal filters, can be expressed in the simple semianalytic form

$$\vec{f}_k = \sum_p C_{kp}\mathbf{V}^{-1}\vec{s}^{(p)}, \quad k \in [C], \tag{14}$$

where the mean traces $\vec{s}^{(p)}$ and correlation matrix $\mathbf{V} = \sum_p \mathbf{\Sigma}^{(p)}$ can both be empirically estimated from data under the known initial state p,

$$\begin{aligned} \vec{s}^{(p)} &\simeq \frac{1}{N_{\text{train}}}\sum_{n=1}^{N_{\text{train}}} \vec{x}_{(n)}^{(p)}, \\ \Sigma^{(p)} &\simeq \frac{1}{N_{\text{train}}}\sum_{n=1}^{N_{\text{train}}} \vec{x}_{(n)}^{(p)}\vec{x}_{(n)}^{(p)T} - \vec{s}^{(p)}\vec{s}^{(p)T}, \end{aligned} \tag{15}$$

while the coefficients C_{kp} can also be shown to depend only on $\vec{s}^{(p)}$ and $\mathbf{V}$ (see Appendix D for full details). Furthermore, the C filters are not all independent; they can be shown to satisfy the constraint (see Appendix D)

$$\sum_{k=1}^{C}\vec{f}_k = \vec{0}, \tag{16}$$

where $\vec{0} \in \mathbb{R}^{N_O N_T}$ is the null vector. This powerful constraint, which holds regardless of the statistics of the noise $\vec{\zeta}$, implies that only $C-1$ of the C filters need to be learned from training data.

A. TPP performance under Gaussian white noise in comparison with standard FGDA

We can now analyze the case most often assumed in cQED: that the dominant noise source in heterodyne records I, Q is stationary Gaussian white noise (independent of the undetermined state), an assumption under which matched filters are optimal for binary classification. Engineering of cQED measurement chains is geared towards approaching this limit, by (1) developing large-bandwidth, high-dynamic-range amplifiers that operate with fast response times and minimal nonlinear effects even at high gain and high input signal powers [48–53], (2) increasing qubit T_1 and tolerance to strong cavity drives to reduce transitions during $\mathcal{T}_{\text{meas}}$ [3], and (3) controlling technical noise sources such as electronic white noise from classical cryo-HEMT amplifiers and room-temperature electronics.

In this relevant limit, the correlation matrix $\mathbf{V}$ of Eq. (5) becomes proportional to the identity matrix, and the resulting TPP-learned filters depend chiefly only on the mean traces $\vec{s}^{(p)}$. For any $C = 2$ state classification task, for example $p \in \{e, g\}$ qubit readout, we can show that $C_{ke} = -C_{kg}$, which reduces $\vec{f}_k$ *exactly* to a standard binary matched filter. Remarkably, the TPP-learned optimal filters in the Gaussian white noise approximation then provide a semianalytically calculable generalization of matched filters to C states.

We can now analyze the multistate classification performance enabled by these TPP-learned optimal filters in comparison with the standard FGDA approach. To guarantee dispersive qubit readout data that are subject only to

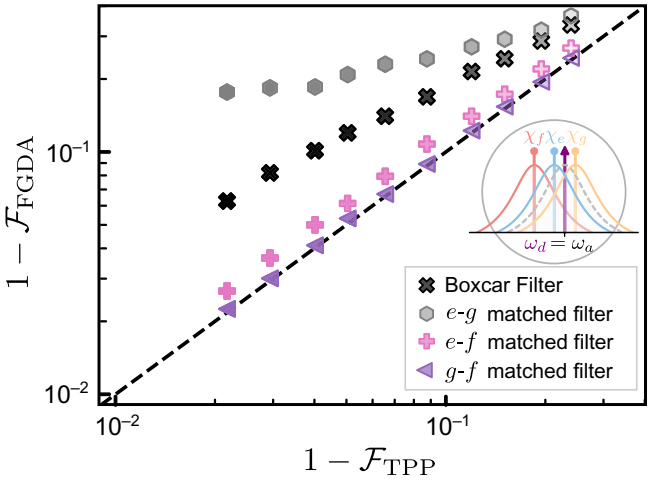


FIG. 2. Multistate ($C = 3$) classification performance of the TPP versus FGDA under Gaussian white noise conditions. We consider dispersive qubit readout to distinguish states $p \in \{e, g, f\}$ as a function of measurement power. For a transmon $\chi_p/\kappa \in \{-\chi, \chi, -3\chi\}$, $\chi/\kappa = 0.195$, and $\kappa/2\pi = 1.54$ MHz. More-opaque markers indicate higher measurement tone amplitudes. The inset shows induced dispersive shifts for each state (not to scale). Standard FGDA is performed with one of three MFs corresponding to each distinct state pair, as well as a boxcar filter. TPP filters are also followed by a Gaussian discriminator for an equivalent comparison. Only one of the binary MFs allows the FGDA to approach the TPP in performance, while all other filters chosen yield a worse performance.

white noise, we use a theoretical simulation of Eq. (6) to generate measured heterodyne records for C qubit states, under the following assumptions: (1) all qubit state transitions are ignored, (2) any additional classical noise sources in the measurement chain are ignored, and (3) therefore direct readout of the cavity can be considered instead of the use of a quantum amplifier and the potential quantum noise added by it. We take the cavity measurement tone to be applied for a subset of the total $\mathcal{T}_{\mathrm{meas}}$, namely, for $[\mathcal{T}_{\mathrm{on}}, \mathcal{T}_{\mathrm{off}}]$, and to be coincident with the cavity center frequency so that $\Delta_{da} = 0$, which is usual for transmon readout (for full details, see Appendix B 1). Other system parameters can be found in the caption for Fig. 2.

The TPP can be used to generate optimal filters, and hence perform classification, for arbitrary C; for examples of calculated filters, see Fig. 10 in Appendix D. For concreteness, here we analyze the classification performance enabled by the TPP to distinguish $C = 3$ states $p \in \{e, g, f\}$. Our choice of resonantly driving the readout cavity means the sign of cavity dispersive shifts for transmon states e and f is the same, and is opposite that for g, making them harder to distinguish (see also the inset in Fig. 2). The specific details of the readout scheme do not change the TPP learning procedure.

For this three-state classification task, a unique filter choice for the FGDA is not known. While certain approaches for constructing filters have been attempted [54], boxcar filtering is still commonly used. Another approach might be to use a matched filter that optimizes distinction of just one pair of states. There are three such filters in total: for discrimination of e-g states as defined in Eq. (11), as well as analogously defined filters for e-f and g-f states.

In Fig. 2, we show classification infidelities $1 - \mathcal{F}$ calculated for datasets with increasing measurement tone amplitude (more opaque markers) with both the optimal TPP filter and the FGDA with the four aforementioned filter choices. We emphasize again that these datasets are generated via simplified theoretical simulations guaranteeing white noise conditions, in particular ignoring any nonidealities associated with strong readout drives; under these conditions, classification performance improves steadily with increasing measurement tone amplitude, as shown. Even in this regime, we clearly observe that the FGDA infidelities for most filter choices are worse than the TPP infidelities. Interestingly, the poorest performer is not the boxcar filter; rather, it is the e-g filter, which would be optimal if we were distinguishing only $\{e, g\}$ states, which yields the worst performance. This is because the e-g filter is completely unaware of the f state: it attempts to best discriminate e and g, but in doing so, it substantially confuses e and f states, which are already the hardest to distinguish. The e-f filter corrects this major problem and hence performs better, but does not discriminate e and g states as well as the e-g filter would. Because of the specific driving conditions and phases, the g-f filter unwittingly does a good job at addressing both these problems, yielding the best performance. Nevertheless, it can only match the performance of the TPP.

This trial-and-error approach relies on knowledge of optimal matched filtering from binary classification, but clearly cannot be optimal for $C > 2$: none of the filter choices are informed by the statistical properties of measured data for *all* C classes to be distinguished. Alternative approaches, such as use of multiple classifiers with up to $C - 1$ independent filters (for an equivalent resource cost to the TPP) can account for all classes, but as we show in Appendix D 7, they do not outperform the TPP, and also exhibit a dependence on readout conditions. In either case, the brute-force determination of pairwise matched filters scales at least with the number of distinct state pairs, which grows quadratically with C; this is before one even accounts for fine-tuning of filter coefficients (analogous to learning C_{kp} in the TPP approach). In contrast, the TPP approach provides a simple automated scheme to learn optimal filters, takes data for readout of all classes into account, is model-free and thus applicable to arbitrary readout conditions, and scales only linearly with the task dimension set by C.

However, the true strength of TPP learning arises when noise in measured heterodyne records no longer satisfies the additive Gaussian white noise assumption, which may arise if any of conditions (1)–(3) for qubit measurement chains listed earlier are not met. Departures from this ideal scenario are widely prevalent in cQED, and will be apparent in experimental results presented in the following section. Throughout the rest of this paper, we show how the trainability of the TPP approach enables it to learn filters tailored to these more general noise conditions and consequently outperform the standard FGDA based on binary matched filters.

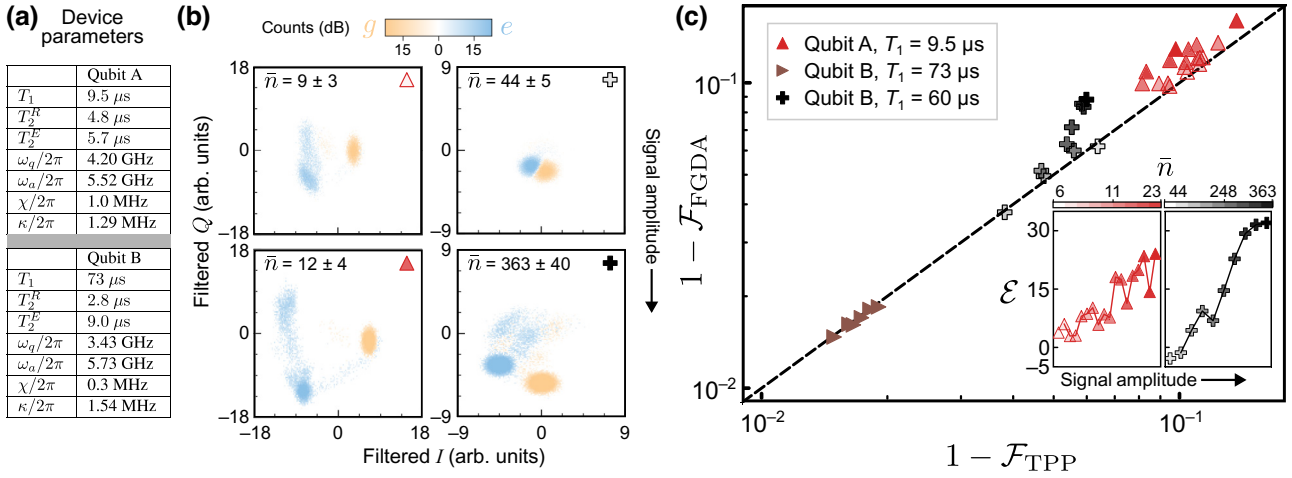
IV. TPP LEARNING FOR REAL QUBITS

A. Experimental results

To demonstrate how the general learning capabilities of the TPP approach can aid qubit state classification in a practical setting, we now apply it to the readout of finite-lifetime qubits in an experimental cQED measurement chain. The essential components of the measurement chain are as depicted schematically in Fig. 1 and described by Eq. (6). The actual circuit diagram is shown in Fig. 8 in Appendix A, and important parameters characterizing the measurement chain components are summarized in Fig. 3(a).

We consider two distinct cavity systems for the dispersive readout of distinct single qubits A and B to discriminate states $p \in \{e, g\}$. For *lossless* qubits that are read out dispersively for a fixed measurement time $\mathcal{T}_{\text{meas}}$, the ratio χ/κ determines the *theoretical* maximum readout fidelity; in particular, an optimal value of this ratio is known under these ideal conditions [42]. However, experimental considerations mean that operating parameters must be designed with several other factors in mind. At high χ/κ ratios with modest or higher κ, for large κ with modest χ/κ ratios, and especially when both are true, the experiment is sensitive to dephasing from the thermal occupation of the readout resonator at a rate proportional to $\bar{n}\kappa$ [55]. This can be quite limiting to the T_2 dephasing time of the qubit if the readout resonator is strongly coupled to the environment and/or the environment has appreciable average thermal photon occupation $\bar{n}$. In the opposite, low-χ/κ limit, the qubit is shielded from thermal dephasing, but readout becomes very difficult as the rate at which one learns about the qubit state from a steady-state coherent drive is proportional to χ/κ [42]. In this experiment, the lower-than-usual $\chi/\kappa \approx 0.2$ in qubit B represents a compromise between these two limits, while also enabling the high-fidelity discrimination of multiple excited states of the transmon (see Fig. 7 in Appendix A).

Each readout cavity is driven in reflection, and its output signal is amplified also in reflection with use of a Josephson parametric amplifier (JPA). We use the latest iteration of strongly pumped and weakly nonlinear JPAs [53], boasting a superior dynamic range. Such JPAs operate well below saturation even at signal powers that correspond to more than 100 photons, enabling us to probe qubit readout at high measurement powers. By choosing a signal frequency of exactly half the pump frequency, we can operate the JPA in phase-sensitive mode. We can also operate the amplifier in phase-preserving mode if we detune the signal



	Qubit A
T_1	9.5 μs
T_2^R	4.8 μs
T_2^E	5.7 μs
$\omega_q/2\pi$	4.20 GHz
$\omega_a/2\pi$	5.52 GHz
$\chi/2\pi$	1.0 MHz
$\kappa/2\pi$	1.29 MHz
	Qubit B
T_1	73 μs
T_2^R	2.8 μs
T_2^E	9.0 μs
$\omega_q/2\pi$	3.43 GHz
$\omega_a/2\pi$	5.73 GHz
$\chi/2\pi$	0.3 MHz
$\kappa/2\pi$	1.54 MHz

FIG. 3. Classification performance of the TPP versus FGDA for readout of real qubits. (a) Parameters of various dispersive qubit-cavity systems used for gathering readout data. Coherence measurements are subject to 10% variation over time. (b) Representative qubit readout histograms under boxcar filtering as a function of measurement signal amplitude. (c) Readout data for three dispersive qubit-cavity systems are analyzed and the resulting classification infidelities for binary ($C = 2$) state classification are plotted against each other. The dashed line marks $1 - \mathcal{F}_{\text{FGDA}} = 1 - \mathcal{F}_{\text{TPP}}$. For datasets with variable shading of markers (red and black), more-opaque markers indicate higher measurement tone amplitudes, with the corresponding resonator photon number $\bar{n}$ indicated via color bars. The inset shows percentage fewer errors $\mathcal{E}$ computed for the indicated datasets with increasing input signal amplitude.

from half the pump frequency by more than the spectral width of the pulse. Several filters are used to reject the strong JPA pump tone required to enable this operation. Circulators are used to route the output signals away from the input signals and to isolate the qubit from amplified noise.

In ideal circumstances, the use of stronger measurement tones should increase the classification fidelity for qubit readout, as shown via simplified theoretical simulations in Fig. 2. In practice, however, higher measurement powers are known to be associated with a variety of complex dynamical effects that can limit fidelity. Perhaps the most common observation is enhanced qubit $e \to g$ decay under strong driving (referred to as the T_1 versus $\bar{n}$ problem). The relative accessibility of higher excited states in transmon qubits means that at strong enough driving, general multilevel transitions to these higher levels can also be observed. There have also been predictions of chaotic dynamics and ionization [14,56] at certain readout resonator occupation levels, as well as complex dynamics due to qubit-induced resonator nonlinearities [57]. The theoretical understanding of these effects and their modeling via an SME analogous to Eq. (6) is an ongoing challenge.

In our experiments, we perform readout across this domain using two different qubits. For qubit A, we simultaneously vary both the pulse amplitude and the pulse duration ($\mathcal{T}_{\text{off}} - \mathcal{T}_{\text{on}}$), the latter from 300 to 1150 ns, to together obtain roughly 9 ± 3 to 18 ± 5 photons in the cavity in the steady state. For qubit B's phase-preserving dataset, the measurement pulse durations vary independently from 500 to 900 ns and the measurement amplitudes are adjusted to drive roughly 44 ± 5 to 363 ± 40 photons in the cavity in the steady state; the significantly larger photon number is tolerated due to the low qubit B χ/κ. For the shortest pulse duration and lowest pulse amplitude, this corresponds to just enough discriminating power to separate the measured distributions for the two states by approximately their width in a boxcar-filtered I-Q plane (namely, without the use of an empirical MF). An example of the individual readout histograms for qubits initialized in states $p \in \{e, g\}$ at this lowest measurement tone power is shown in Fig. 3(b). Qubit B's phase-sensitive dataset was recorded with a pulse time of 800 ns with a shaped pulse to shorten the effect of the cavity ring-up time, similarly to what was done in the work reported in Ref. [58].

At the highest measurement powers, we are able to populate the readout cavity with hundreds of photons, calibrated by our observing the frequency shift of the qubit drive frequency versus the occupation of the readout resonator. At these powers, extreme higher-state transitions become visible during the readout pulse [9]; an example is shown in Fig. 3(b) (see also Fig. 7 in Appendix A). There is also a notable elliptical distortion in the high-amplitude data, particularly for qubit A. We suspect that this is due to the short duration of the pulses and the inclusion of the cavity ring-up and ring-down in the integration, since the simple boxcar filter used to integrate the histograms in Fig. 3(b) does not rotate with the signal mean.

For such complex regimes where no simple model of the dynamics exists, the construction of an optimal filter is not known; this hence serves as an ideal testing ground for the TPP approach to qubit state classification. We compute the infidelities of binary classification using both the TPP scheme and an FGDA using the standard MF [Eq. (11)] under a variety of readout conditions, plotting the results against each other in Fig. 3(c).

The highest fidelity achieved with both schemes is obtained for qubit B under conditions where its T_1 time is longest. This dataset was collected at a fixed, moderate measurement power; the different points correspond to a rolling of the relative JPA pump and measurement tone phase that determines the amplified quadrature under phase-sensitive operation. The dashed line marks equal classification infidelities, so any datasets above this line yield a higher classification *infidelity* with the FGDA than with the TPP. Here we see that both schemes exhibit very similar performance levels.

The other two datasets are obtained for readout under varying measurement powers. The depth of shading of the markers indicates the strength of measurement drives: the more opaque the marker, the greater the measurement power. We first note that the classification fidelity does not uniformly increase with signal amplitude in experiments; this is in contrast to the simplified theoretical simulations in Sec. III A, and is expected due to the aforementioned dynamical effects exhibited in real qubit readout at higher readout powers (ignored in Fig. 2).

For lower measurement powers, we see that the performances of the TPP and the FGDA are once again comparable. However, a very clear trend emerges: for greater measurement powers—where measurement dynamics become much more complex as demonstrated in Fig. 3(b)—the TPP generally outperforms the FGDA. To more precisely quantify the difference in performance between the TPP and the FGDA, we introduce the metric $\mathcal{E}$,

$$\mathcal{E} = \left(\frac{\mathcal{F}_{\text{TPP}} - \mathcal{F}_{\text{FGDA}}}{1 - \mathcal{F}_{\text{FGDA}}} \right) \times 100, \tag{17}$$

which essentially asks: "what percentage fewer errors does the TPP make when compared with the FGDA?" We plot $\mathcal{E}$ in the inset in Fig. 3(c) for the two qubit readout experiments where the input power is varied. We see clearly that with increasing power, the TPP can significantly outperform the FGDA scheme, committing as many as 30% fewer errors in the experiments considered. In certain cases where the FGDA predicts a reduction in classification fidelity with increasing readout power, the TPP's

learning advantage can even enable a qualitatively different trend, instead boosting classification performance with increasing readout power (for details, see Appendix E 1).

Our results demonstrate that the TPP approach can be successfully applied to real qubit readout across a broad spectrum of measurement conditions. Furthermore, the TPP can even outperform the standard FGDA in certain relevant regimes, such as for high-power readout. While the TPP can thus be applied as a model-free learning tool, we are also interested in understanding the principles that enable the TPP to outperform standard approaches using an MF. Uncovering these principles can help identify the types of classification tasks where TPP learning is essential. Our interpretation of TPP learning as optimal filtering proves to be a useful tool in this vein.

B. Adaptation of TPP-learned filters under strong measurement tones

For visualization, we analyze only filters $\vec{f}_k \in \mathbb{R}^{N_T}$ for I-quadrature data; the complete vector $\vec{\boldsymbol{f}}_k$ includes filters for all N_O observables. Recall that for a C state classification task, the TPP learns C filters; however, the sum of filters is constrained by Eq. (16), so $C-1$ filters are sufficient to describe the TPP's learning capabilities. In Fig. 4(a), we first consider filters learned by the TPP for a $C=2$ classification task for select experimental datasets from Fig. 3 obtained under a low measurement power and a high measurement power. It therefore suffices to analyze just $\vec{f}_1$, the first filter for the I quadrature, as a function of measurement power. The black curves represent filters learned under the assumption of Gaussian white noise; recall that for this binary case, these filters are exactly the standard MF. The gray curves, in contrast, represent filters learned by the TPP for arbitrary noise conditions, obtained by our solving Eq. (2). At a low measurement tone amplitude (less opaque marker), the general TPP filter appears very similar to the TPP filter under white noise. As the measurement tone amplitude is increased, however, the TPP-learned filter under arbitrary noise can deviate substantially from the TPP filter under white noise. This is accompanied by a marked difference in performance, as observed in Fig. 3(c).

Crucially, the generalization of matched filters provided by TPP learning as discussed in Sec. III A enables a similar comparison for classification tasks for an arbitrary number of states. We show learned filters for $C=3$ state classification of $p \in \{e,g,f\}$ in Fig. 4(b), again for a low measurement power and a high measurement power. It is now sufficient to consider any two of three distinct I-quadrature filters; here we choose $\vec{f}_1$ and $\vec{f}_3$. Once more, the general TPP filters begin to deviate significantly from TPP filters under the white noise assumption at high powers. Most importantly, these filters provide an increase in three-state classification fidelity relative to the FGDA scheme (for brevity, full results are provided in Appendix E 2).

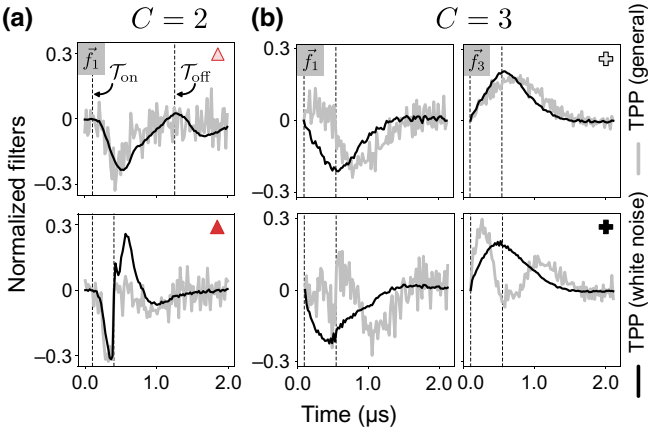


FIG. 4. Adaptation of TPP-learned filters with increasing measurement tone amplitude and evolving noise conditions. Black curves represent normalized TPP filters under the white noise assumption; for binary state classification, these are identical to standard matched filters. Gray curves represent general TPP filters with no assumptions on noise statistics. (a) Filter $\vec{f}_1$ for binary ($C=2$) classification and (b) filters $\vec{f}_1$ and $\vec{f}_3$ for $C=3$ state classification. In both cases, at lower amplitudes, the general TPP filter temporal profile closely matches that of the TPP filter under the assumption of white noise. However, for lower measurement amplitudes, a marked difference between the white noise TPP filter and the general TPP filter is observed.

Clearly, the precise form of filters learned by the TPP to outperform white noise filters must be influenced by some physical phenomena that arise at high measurement powers. However, the TPP is not provided with any physical description for such phenomena, which is part of its model-free appeal. What then is the mechanism through which the TPP can learn about such phenomena to compute optimal filters? The answer lies explicitly in Eq. (14): TPP-learned filters are sensitive to noise correlations in data via $\mathbf{V}$. Using simulations of measurement chains where the noise structure of quantum measurement data can be precisely controlled, we show that the noise structure can strongly deviate from white noise conditions under practical settings. Crucially, the TPP can adapt to these changes, whereas the MF cannot.

V. TPP LEARNING: SIMULATION RESULTS

As discussed in Sec. III, the TPP weights and hence optimal filters depend on mean traces, but are also cognizant of—and can learn from—the noise structure of measured data via the temporal correlation matrix $\mathbf{V}$. This is in stark contrast to the use of a matched filter.

Crucially, data obtained from *quantum* systems can exhibit temporal correlations that have a quantum-mechanical origin. In what follows, we demonstrate the ability of the TPP to learn these quantum correlations, using simulations of two experimental setups where such

quantum noise sources arise naturally: (1) readout using phase-preserving quantum amplifiers with a finite bandwidth, so that the amplifier-added noise (demanded by quantum mechanics) has a nonzero correlation time, and (2) readout of finite-lifetime qubits with multilevel transitions (quantum jumps).

A. Correlated quantum noise added by finite-bandwidth phase-preserving quantum amplifiers

Quantum-limited amplifiers are a mainstay of measurement chains in cQED, are are needed to overcome the added classical noise of following HEMTs. Phase-preserving quantum amplifiers are necessitated by quantum mechanics to add a minimum amount of noise to the incoming cavity signal being processed. The correlation time of this added quantum noise is determined by the dynamics of the amplifier itself, namely, its active linewidth reduced by antidamping necessary for gain. For finite-bandwidth amplifiers operating at large-enough gains, this can lead to the addition of quantum noise with nonzero correlation time in measured heterodyne data.

To simulate qubit readout in these circumstances, we consider a quantum measurement chain described by Eq. (6) now consisting of a qubit-cavity-amplifier setup. $\mathcal{L}_{\rm meas}$ then describes the readout of a nondegenerate (i.e., two-mode) parametric amplifier and its nonreciprocal coupling to the cavity used to monitor the qubit. We ignore qubit state transitions, so $\mathcal{L}_{\rm envt}$ describes only losses via unmonitored ports of the cavity and amplifier. Full details of the simulated SME are included in Appendix B 2.

We must consider added classical noise in the measurement chain, as this is what demands the use of a quantum amplifier in the first place. We take the added classical noise to be purely white noise, $\xi^{\rm cl}(t_i) = \sqrt{\bar{n}_{\rm cl}}\frac{dW}{dt}(t_i)$, with noise power $\bar{n}_{\rm cl} = 30$, parameterized as usual in "photon number" units; these assumptions on the noise structure and power are taken from standard cQED experiments, including our own. The obtained heterodyne measurement records, Eqs. (8a) and (8b), then contain two dominant noise sources: (1) excess classical white noise and (2) quantum noise added by the amplifier, contained once again in quantum trajectories $\langle \hat{X}^{(\sigma)}(t)\rangle_c$ and $\langle \hat{P}^{(\sigma)}(t)\rangle_c$.

We restrict ourselves to binary classification of states $|e\rangle$ and $|g\rangle$; here the matched filtering scheme is unambiguously defined and serves as a concrete benchmark for comparison with the TPP approach. In Fig. 5, we compare infidelities calculated with use of the FGDA and TPP approaches for three different values of amplifier transmission gain $\mathcal{G}_{\rm tr}$ and as a function of the coherent input tone power: darker markers correspond to readout with stronger input tones.

To understand how correlations in the measured data depend on the varying amplifier gain, we introduce the noise power spectral density (PSD) of the data (here, the

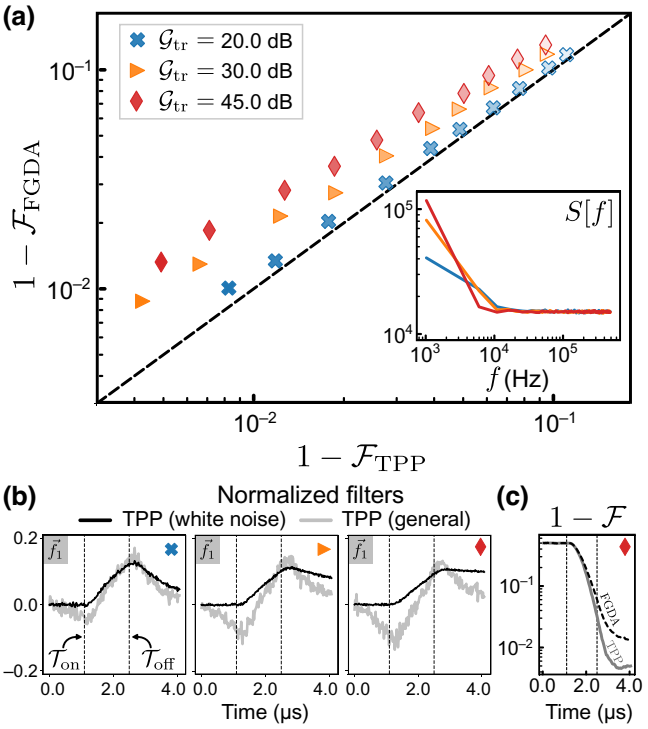


FIG. 5. Classification performance of the TPP versus FGDA on a simulated dataset for readout via a phase-preserving quantum amplifier. (a) Classification infidelities for various amplifier transmission gains $\mathcal{G}_{\rm tr}$ as a function of measurement signal amplitude (more-opaque markers represent higher amplitudes). The ratio of the bare amplifier linewidth to the cavity mode linewidth is $\gamma/\kappa = 5$. Noise PSD is shown in the inset for the different operating gains (for a linear amplifier, this is independent of the measurement signal amplitude). (b) Learned filters under the white noise assumption (black) and general noise conditions (gray) for representative datasets of each value of $\mathcal{G}_{\rm tr}$. (c) Classification infidelities as a function of total time t. The measurement tone is ON only between the two dashed vertical lines.

I quadrature) for state $|p\rangle$,

$$S^{(p)}[f\,] \approx \sum_{j>k}^{N_{\rm T}} e^{-i2\pi f\,\tau_{jk}}\,\mathbf{\Sigma}^{(p)}_{jk}, \tag{18}$$

where $\tau_{jk} = \Delta t(j-k)$. The PSD is simply the Fourier transform of the noise autocorrelation function (by the Wiener-Khinchin theorem). Through $\mathbf{V}$, the TPP learns from these correlations when optimizing filters. The noise PSD is plotted in the inset in Fig. 5; for the current readout task, it is independent of p. With increasing gain, the PSD deviates from the flat spectrum representative of white noise to a spectrum that peaks at low frequencies, indicative of an extended correlation time. The observations also emphasize that noise added by the quantum amplifier dominates over heterodyne measurement noise ξ, as well as excess classical noise $\xi^{\rm cl}$.

For the lowest amplifier gain considered, we see that the FGDA classification performance and the TPP classification performance are quite close to each other. However, with increasing gain, the FGDA infidelity is substantially higher, up to an order of magnitude worse for the largest gain considered here. This TPP performance advantage is enabled by optimized filters, as shown in Fig. 5(b). The measurement tone is ON only between the two dashed vertical lines. The curves in black represent white noise filters, exactly equal to the MF in this binary case. Note that these filters also change with gain: the amplifier response time increases at higher gains, so the mean traces and hence the MF derived from these traces exhibit much slower rise and fall times. The general TPP filter is similar to the MF at low gains, but becomes markedly distinct at higher gains.

Interestingly, one such change is that at high gains the general TPP filter becomes nonzero even before the turning on of the measurement signal (the first vertical dashed line). This appears odd at first sight, since there must not be any information that could enable state classification before a measurement tone probes the cavity used for dispersive qubit measurement. To validate this, in Fig. 5(d) we plot $1-\mathcal{F}$ calculated for an increasing length of measured data, $t \in [0, \mathcal{T}_{\text{meas}}]$. We clearly see that for $t < \mathcal{T}_{\text{on}}$, both the TPP and the FGDA cannot distinguish the states, as must be the case. The nonzero segment of the general TPP filter before $\mathcal{T}_{\text{on}}$ instead accounts for noise correlations. In particular, because of the long correlation time of noise added by the quantum amplifier, noise in data beyond $\mathcal{T}_{\text{on}}$ is correlated with noise from $t < \mathcal{T}_{\text{on}}$. The general TPP filter is aware of these correlations that the standard MF is completely oblivious to, and by accounting for them, it improves classification performance.

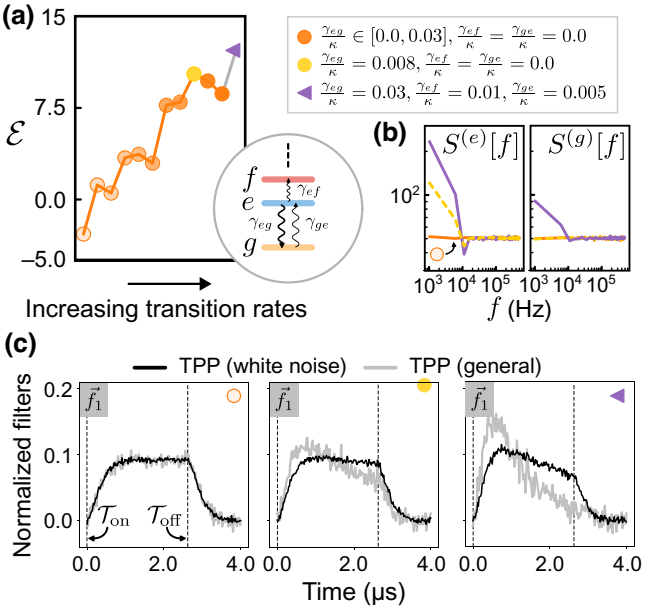


FIG. 6. Classification performance of the TPP versus FGDA on a simulated dataset for readout of a qubit experiencing multilevel transitions. (a) $\mathcal{E}$ as a function of increasing transition rate (more-opaque markers). The schematic in the inset shows the transmon levels and nonzero transition rates considered. (b) Noise PSD $S^{(p)}[f]$ for three representative datasets in the inset in (a), indicating deviation from flat (white noise) as measurement data include more transitions. (c) TPP-learned filters (gray) compared with matched filters (black) for representative datasets, showing adaptation with increasing transition rates.

B. Correlated quantum noise due to multilevel transitions

A transmon is a multilevel artificial atom, as described by Eq. (7); as a result, it is possible to excite levels beyond the typical two-level computational subspace of e and g states. Such transitions manifest themselves as stochastic quantum jumps in quantum measurement data and are an important source of error in readout.

To model measurement under such conditions, we now consider the dispersive heterodyne readout of a finite-lifetime transmon with possible occupied levels $\{e, g, f\}$. We further allow only a subset of all possible allowed transitions between these levels, and with static rates: $|e\rangle \to |g\rangle$ at rate γ_{eg}, the reverse $|g\rangle \to |e\rangle$ at rate γ_{ge}, and $|e\rangle \to |f\rangle$ at rate γ_{ef} (see the inset in Fig. 6). The transitions are described by the superoperator $\mathcal{L}_{\text{envt}}$, while $\mathcal{L}_{\text{meas}}$ describes the measurement tone incident on the cavity and the heterodyne measurement superoperator for the same; for full details, see Appendix B 3.

For simplicity, we now further ignore excess classical noise added by the measurement chain, dropping terms $\xi_I^{\text{cl}}(t_i), \xi_Q^{\text{cl}}(t_i)$. As a result, the measurement records obtained, Eqs. (8a) and (8b), contain only two noise sources: white heterodyne measurement noise and quantum noise due to qubit state transitions imprinted on the emanated cavity field, contained in quantum trajectories of cavity quadratures $\langle \hat{X}^{(\sigma)}(t)\rangle_c$ and $\langle \hat{P}^{(\sigma)}(t)\rangle_c$. We then generate simulated datasets by integrating the resulting full SME, Eq. (6), for different values of transition rates, and consider the task of binary classification of states $p \in \{e, g\}$.

We compare the performance of a trained TPP against that of an FGDA with an empirical MF using the metric $\mathcal{E}$ in Fig. 6(a) with varying transition rates. The noise PSD is plotted in Fig. 6(b) for representative datasets. In the absence of any transitions (lightest orange), $S^{(p)}[f]$ is flat at all frequencies, regardless of the initially prepared state p. This is because the measured data have only heterodyne white noise. With an increase in γ_{eg}, we note that $S^{(e)}[f]$ deviates from the white noise spectrum, attaining a peak at low frequencies. In contrast, $S^{(g)}[f]$ remains unchanged as trajectories for initial states $|g\rangle$ undergo no transitions. In the most complex case, where we allow for

all considered transitions, $S^{(g)}[f]$ also starts to demonstrate deviation from the white noise spectrum.

From readout datasets with no transitions to readout data with increasing transition rates, we note a small but clear improvement in classification performance with the trained TPP in comparison with the FGDA. That the TPP is able to learn information in the presence of transitions that evades the MF is clear when we compare the two sets of filters in Fig. 6(c). As the transition rates increase, the MF undergoes modifications due to the changes to the means of heterodyne records. However, the TPP is sensitive to changes beyond means—in the correlations of measured data—and increasingly learns a distinct filter with sharply decaying features. The utility of similar exponential linear filters for finite-lifetime qubits was the subject of earlier analytic work [21]. The TPP approach generalizes the ability to learn such filters in the presence of arbitrary transition rates and measurement tones, and for multistate classification.

One may note that in the absence of any multilevel transitions [Fig. 6(a), first data point] the FGDA appears to outperform the TPP ($\mathcal{E} < 0$); given the results presented in Sec. III A, this may seem odd, as here the measurement noise is exactly Gaussian white noise, so the TPP filter reduces exactly to the MF used in the standard FGDA. The important distinction is that, unlike in Sec. III A, here we are deploying the *general* TPP, which makes no *a priori* assumptions about noise characteristics. In the special case where the noise is Gaussian white noise, the MF is already cognizant of the correct noise statistics, while the TPP must learn them via training, leading to a slight underperformance that is alleviated as the size of the training dataset is increased (see also Appendix C). Of course, this freedom is precisely what enables the TPP to learn more efficiently when the noise characteristics are not simply Gaussian and white, for example, under increasing multilevel transitions. There, the TPP shows an improvement relative to the standard FGDA *in spite* of having to learn the new noise statistics from training data. For these more complex noise conditions, the standard MF is now suboptimal, and the FGDA performance suffers as a result.

Finally, we emphasize that the simplified transition model considered here is chosen to highlight the ability of the TPP to learn quantum noise associated with quantum jumps under controlled noise conditions, where no other nontrivial noise sources (classical or quantum) exist. The TPP approach to learning is model-free, and its ability to learn in more general noise settings is demonstrated by its adaptation to real qubit readout in Sec. IV.

VI. DISCUSSION AND OUTLOOK

In this paper we have demonstrated a machine-learning approach to classification of an arbitrary number of states using temporal data obtained from quantum measurement chains. While we have focused on the task of dispersive readout of multilevel transmons, the TPP approach applies broadly to quantum systems, and more generally physical systems, monitored over time. Our results show that the TPP framework for processing quantum measurement data reduces to standard approaches based on matched filtering in the precise regimes of validity of the latter. However, the TPP can adapt to more general readout scenarios to significantly outperform matched filtering schemes. We show this improvement for the TPP trained on real qubit readout data to confirm the practical utility of our scheme.

Rather than treating the TPP as a black box, in our work we clarify the learning mechanism that enables the TPP to outperform matched filtering schemes. First, we develop a heuristic interpretation of the TPP mapping as one of applying temporal filters to measured data. TPP learning then amounts to learning optimal filters. Deconstructing the learning scheme, we find the TPP performance advantage is enabled by its ability to learn optimal filters by accounting for noise *correlations* in temporal data. When this noise is purely white noise, the TPP approach provides a generalization of matched filtering to an arbitrary number of states.

Crucially, we find that the TPP can efficiently learn from correlations not just due to classical signals, or in principle due to quantum noise in theory, but also from practical systems where most of the noise is quantum in origin. In addition to real qubit readout, using theoretical simulations where the strength of quantum noise sources can be tuned precisely, such as noise due to multilevel transitions or the added noise of phase-preserving quantum amplifiers, we clearly demonstrate that the TPP can learn from quantum noise correlations to outperform standard matched filtering. Furthermore, our precise identification of quantum correlations as a harnessable resource can help guide future machine-learning approaches to quantum signal processing.

The TPP approach, anchored by its connection to standard matched filtering, with demonstrated advantages for real qubit readout under complex readout conditions, and feasibility for FPGA implementations (to be demonstrated in future work), is ideal for integration with cQED measurement chains for the next step in readout optimization. Furthermore, the TPP's generality and ability to efficiently learn from data could pave the way for an even broader class of applications. An important potential use is as a postprocessor of quantum measurement data for quantum machine learning. With the use of general quantum machines for information processing, the optimal means to extract data from their measurements may not always be known. We believe the TPP is ideally suited to uncover the optimal linear postprocessing step, through training that could be incorporated as part of the optimization of the quantum machine. This is because the existence of an exact analytic form for the optimal trained TPP weights

eliminates the need for multiple training epochs, batchwise evaluations, or gradient computations, so training the TPP adds minimal complexity to the optimization of an already complex quantum measurement chain, in stark contrast to the substantial overhead of training a neural network used as a postprocessor. Finally, optimal state estimation is essential for control applications. The trainable TPP can form part of a framework for control applications, such as Kalman filtering for quantum systems.

ACKNOWLEDGMENTS

We thank Leon Bello, Dan Gauthier, and Shyam Shankar for useful discussions. This work was supported by the AFOSR under Grant No. FA9550-20-1-0177, by the Army Research Office under Grant No. W911NF18-1-0144, and by the J. Insley Blair Pyne Fund.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the AFOSR, the Army Research Office, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

APPENDIX A: EXPERIMENTAL SETUP

In this appendix, we show a few more examples of readout I-Q histograms as well as a more detailed circuit diagram for the measurement chain. Shown in Fig. 7, we see two examples of the extremes of the measurement data for readout of qubit B used to generate Fig. 3. Figure 7(a) shows results for a lower-power readout pulse applied for a short time of 300 ns, where the cavity barely has time to reach a steady state before the drive is turned off. Consequently, information from both the ring-up and the ring-down must be integrated to achieve the SNR shown in this figure. Despite this measure, there is still significant infidelity from the lack of separation of the Gaussian signals. In the second case, the displacement voltage is larger, and the pulse is 3 times as long, resulting in significantly increased separation of the Gaussian signals and enabling discrimination of the $|g\rangle$, $|e\rangle$, $|f\rangle$, and $|h\rangle$ states. However, the high powers required induce transitions between these states, resulting in the trails between them as the measurement integrates a mixture of different cavity states at different times.

In Fig. 8, a schematic of the hardware used for the measurements reported in Sec. IV is shown. The measurement setup is fairly standard, with use of single sideband upconversion to send signals into the dilution refrigerator, moving through three stages of attenuation, with 20-dB attenuation at 4 K, 20-dB attenuation at the 100-mK stage, and approximately 45-dB attenuation at the base stage of the refrigerator, with 10 dB of the base-stage attenuation coming from a particularly-well-thermalized copper-body attenuator. The signal interacts with the qubit and cavity system, is routed by two circulation stages to the amplifier, is amplified in reflection, and then is routed once again back through the circulators to the remaining stages of amplification at 4 K and room temperature accordingly. From there it is down-converted by the same local oscillator to 50 MHz, filtered, amplified once more at low frequency, digitized at 1×10^9 samples per second, and

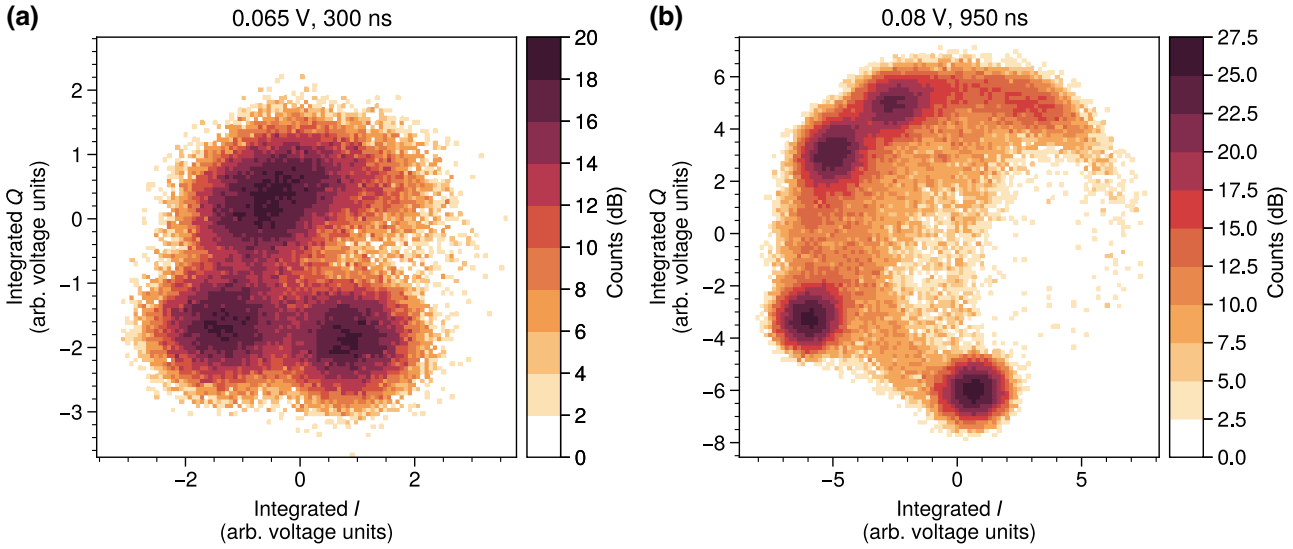


FIG. 7. Comparison between boxcar-integrated I-Q results for (a) a lower-power pulse applied for a short time, corresponding to $\bar{n} = 116$ readout photons, and (b) a higher-power pulse applied for a longer time, corresponding to $\bar{n} = 176$ readout photons. State transitions are visible as "trails" leading between the primary symbols in (b). Counts are shown in logarithmic units to emphasize low-count trails.

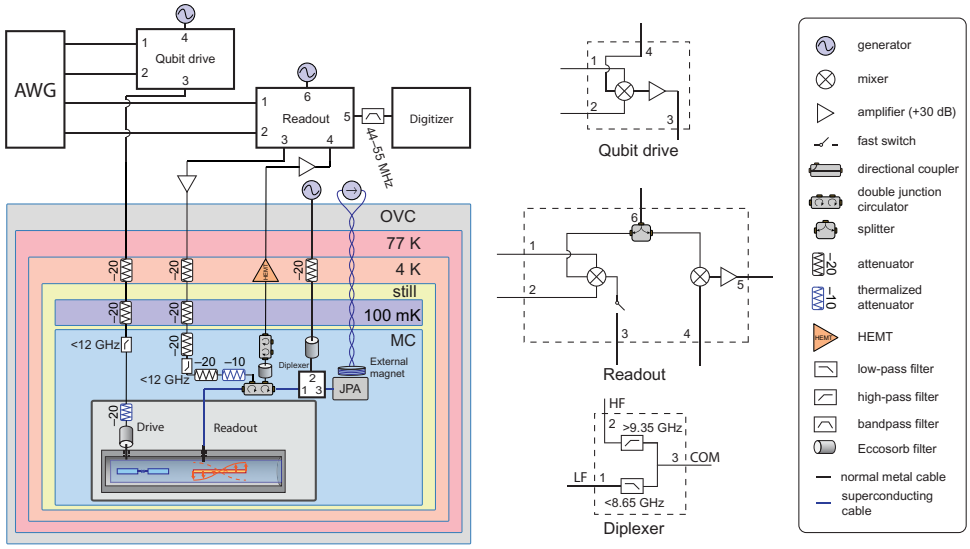


FIG. 8. Up-conversion and down-conversion schematic for drive pulses sent first to the qubit readout resonator, driven in reflection, and then routed to the amplifier and to the HEMT via two circulators. AWG, arbitrary-waveform generator; HF, high frequency; LF, low frequency; MC, mixing chamber; OVC, outer vacuum chamber.

finally demodulated and integrated to produce a readout histogram such as the histograms shown in Fig. 7.

APPENDIX B: SIMULATING HETERODYNE MEASUREMENT RECORDS OBTAINED FROM QUANTUM MEASUREMENT CHAINS FOR DISPERSIVE QUBIT READOUT

In this appendix, we describe the SMEs used to model various quantum measurement chains and generated datasets analyzed in the main text. For convenience we reproduce the general SME of Eq. (6):

$$d\hat{\rho}_c = \mathcal{L}_{\rm sys}\hat{\rho}_c\,dt + \mathcal{L}_{\rm envt}\hat{\rho}_c + \mathcal{L}_{\rm meas}[dW]\hat{\rho}_c. \tag{B1}$$

For all the considered models of quantum measurement chains for the fixed task of dispersive qubit readout, $\mathcal{L}_{\rm sys}$ remains the same, as identified in the main text:

$$\mathcal{L}_{\rm sys}\hat{\rho}_c = -i[\hat{\mathcal{H}}_{\rm disp}, \hat{\rho}_c], \tag{B2}$$

where $\hat{\mathcal{H}}_{\rm disp}$ is the dispersive cQED Hamiltonian for a multilevel artificial atom,

$$\hat{\mathcal{H}}_{\rm disp} \simeq \sum_p \omega_p |p\rangle\langle p| - \Delta_{da}\hat{a}^\dagger\hat{a} + \sum_p \chi_p \hat{a}^\dagger\hat{a}|p\rangle\langle p|. \tag{B3}$$

The superoperators $\mathcal{L}_{\rm envt}$ and $\mathcal{L}_{\rm meas}[dW]$ will depend on the specific model considered.

1. Dispersive readout with no qubit transitions and using a cavity

For qubit readout in the absence of any state transitions, $\mathcal{L}_{\rm envt} \to 0$. As a result, the SME of Eq. (B1) takes a simpler form:

$$d\hat{\rho}_c = \mathcal{L}_{\rm sys}\hat{\rho}_c\,dt + \mathcal{L}_{\rm meas}[dW]\hat{\rho}_c. \tag{B4}$$

Here $\mathcal{L}_{\rm sys}$ is given by Eq. (B2). The superoperator $\mathcal{L}_{\rm meas}$ describes quantum modes in the measurement chain that are used to measure the quantum system of interest. This superoperator can be expressed in the general form

$$\mathcal{L}_{\rm meas}[dW]\hat{\rho}_c = \mathcal{L}_{\rm q}\hat{\rho}_c + \mathcal{S}[dW]\hat{\rho}_c. \tag{B5}$$

Here $\mathcal{L}_{\rm q}$ defines the unconditional dynamics of quantum modes used for measurement; here it takes the explicit form

$$\mathcal{L}_{\rm q}\hat{\rho} = -i[\eta(\hat{a}+\hat{a}^\dagger), \hat{\rho}] + \kappa\mathcal{D}[\hat{a}]\hat{\rho}, \tag{B6}$$

which describes the measurement tone used for cavity readout and the cavity losses due to its monitored port. Importantly, $\mathcal{L}_{\rm q}$ is independent of the qubit sector.

$\mathcal{S}[dW]$ is the stochastic measurement superoperator that describes conditional evolution under continuous heterodyne monitoring:

$$\mathcal{S}[dW]\hat{\rho}_c = \sqrt{\frac{\kappa}{2}}\left(\hat{a}\hat{\rho}_c + \hat{\rho}_c\hat{a}^\dagger - \langle\hat{a}+\hat{a}^\dagger\rangle\hat{\rho}_c\right)dW_I + \sqrt{\frac{\kappa}{2}}\left(-i\hat{a}\hat{\rho}_c + i\hat{\rho}_c\hat{a}^\dagger - \langle -i\hat{a} + i\hat{a}^\dagger\rangle\hat{\rho}_c\right)dW_Q. \tag{B7}$$

These explicit forms of superoperators fully define Eq. (B1) in this regime without qubit transitions. However,

this simplifying assumption can be used to further simplify the form of the SME. In particular, in the absence of transitions, the quantum state of the measurement chain is given by the ansatz

$$\hat{\rho}(t) = |p\rangle\langle p| \otimes \hat{\varrho}_c(t), \tag{B8}$$

where $\hat{\varrho}_c(t)$ is the conditional density matrix defining the quantum state of all quantum modes in the measurement chain *other* than the qubit (namely, the cavity mode). The above implies that the qubit state is completely unchanged during the readout time. The only evolution is in the state of the modes used to read out the qubit, namely, the cavity modes.

By now tracing out the qubit subspace in Eq. (B1), we can obtain an SME for $\hat{\varrho}_c(t)$ alone, under the ansatz of Eq. (B8). The Hamiltonian contribution from the dispersive qubit Hamiltonian yields

$$\begin{aligned}
&\mathrm{tr}_Q\{\hat{\mathcal{H}}_{\mathrm{disp}}|p\rangle\langle p| \otimes \hat{\varrho}_c\} \\
&\quad = \mathrm{tr}_Q\Big\{\sum_j \omega_j\, |j\rangle\langle j|\, p\rangle\langle p| \otimes \hat{\varrho}_c\Big\} \\
&\qquad - \mathrm{tr}_Q\Big\{|p\rangle\langle p| \otimes (\Delta_{da}\hat{a}^\dagger\hat{a}\hat{\varrho}_c)\Big\} \\
&\qquad + \mathrm{tr}_Q\Big\{\sum_j \chi_j \hat{a}^\dagger\hat{a}\, |j\rangle \underbrace{\langle j|p\rangle}_{\delta_{jp}} \langle p| \otimes \hat{\varrho}_c\Big\} \\
&\quad = \omega_p\hat{\varrho}_c - \Delta_{da}\hat{a}^\dagger\hat{a}\hat{\varrho}_c + \chi_p\hat{a}^\dagger\hat{a}\hat{\varrho}_c,
\end{aligned} \tag{B9}$$

and by conjugation

$$\mathrm{tr}_Q\{|p\rangle\langle p| \otimes \hat{\varrho}_c\hat{\mathcal{H}}_{\mathrm{disp}}\} = \hat{\varrho}_c\omega_p - \hat{\varrho}_c\Delta_{da}\hat{a}^\dagger\hat{a} + \hat{\varrho}_c\chi_p\hat{a}^\dagger\hat{a}, \tag{B10}$$

following which we arrive at

$$\begin{aligned}
\mathrm{tr}_Q\{-i[\hat{\mathcal{H}}_{\mathrm{disp}}, \hat{\rho}_c]\} &= -i\left([-\Delta_{da}\hat{a}^\dagger\hat{a}, \hat{\varrho}_c] + [\chi_p\hat{a}^\dagger\hat{a}, \hat{\varrho}_c]\right) \\
&= -i[\left(-\Delta_{da} + \chi_p\right)\hat{a}^\dagger\hat{a}, \hat{\varrho}_c] \\
&\equiv -i[\hat{\mathcal{H}}_{\mathrm{cav}}, \hat{\varrho}_c],
\end{aligned} \tag{B11}$$

where we have defined $\hat{\mathcal{H}}_{\mathrm{cav}}$ as the cavity Hamiltonian alone,

$$\hat{\mathcal{H}}_{\mathrm{cav}} = \left(-\Delta_{da} + \chi_p\right)\hat{a}^\dagger\hat{a} = (\omega_a + \chi_p - \omega_d)\hat{a}^\dagger\hat{a}. \tag{B12}$$

We can perform a similar simplification of terms due to $\mathcal{L}_{\mathrm{meas}}$. For the ansatz in Eq. (B8), we find for $\mathcal{L}_{\mathrm{q}}$,

$$\begin{aligned}
\mathrm{tr}_Q\{\mathcal{L}_{\mathrm{q}}(|p\rangle\langle p| \otimes \hat{\varrho}_c)\} &= \mathrm{tr}_Q\{|p\rangle\langle p| \otimes \mathcal{L}_{\mathrm{q}}\hat{\varrho}_c\} \\
&= \mathrm{tr}_Q\{|p\rangle\langle p|\} \otimes \mathcal{L}_{\mathrm{q}}\hat{\varrho}_c = \mathcal{L}_{\mathrm{q}}\hat{\varrho}_c.
\end{aligned} \tag{B13}$$

As $\mathcal{L}_{\mathrm{q}}$ was independent of the qubit subsector, it remains unchanged following the partial trace over this subsector. The stochastic measurement operator $\mathcal{S}[dW]$ is again independent of the qubit subspace. Hence, our tracing out the qubit sector yields

$$\begin{aligned}
\sqrt{\kappa}\ \mathrm{tr}_Q\{\mathcal{S}[dW]|p\rangle\langle p| \otimes \hat{\varrho}_c\} &= \sqrt{\kappa}\ \mathrm{tr}_Q\{|p\rangle\langle p|\} \otimes \mathcal{S}[dW]\hat{\varrho}_c \\
&= \sqrt{\kappa}\mathcal{S}[dW]\hat{\varrho}_c.
\end{aligned} \tag{B14}$$

The final *cavity-only* SME in the absence of any qubit transitions takes the form

$$d\hat{\varrho}_c = -i[\hat{\mathcal{H}}_{\mathrm{cav}}, \hat{\varrho}_c]dt + \mathcal{L}_{\mathrm{meas}}[dW]\hat{\varrho}_c. \tag{B15}$$

The resulting SME preserves Gaussian states and can thus be solved exactly by a truncated equations of motion approach.

2. Dispersive readout with no qubit transitions and using a quantum-limited amplifier with added noise

As in the previous subsection, in the absence of any state transitions, $\mathcal{L}_{\mathrm{envt}} \to 0$, and the SME of Eq. (B1) takes the simpler form

$$d\hat{\rho}_c = \mathcal{L}_{\mathrm{sys}}\hat{\rho}_c\, dt + \mathcal{L}_{\mathrm{meas}}[dW]\hat{\rho}_c. \tag{B16}$$

Again $\mathcal{L}_{\mathrm{sys}}$ is given by Eq. (B2), and $\mathcal{L}_{\mathrm{meas}}$ takes the form

$$\mathcal{L}_{\mathrm{meas}}[dW]\hat{\rho}_c = \mathcal{L}_{\mathrm{q}}\hat{\rho}_c + \mathcal{S}[dW]\hat{\rho}_c. \tag{B17}$$

$\mathcal{L}_{\mathrm{q}}$ for the unconditional dynamics of quantum modes used for measurement takes the explicit form

$$\mathcal{L}_{\mathrm{q}}\hat{\rho} = -i[\eta(\hat{a} + \hat{a}^\dagger), \hat{\rho}] + \kappa'\mathcal{D}[\hat{a}]\hat{\rho} + \mathcal{L}_c\hat{\rho}_c + \mathcal{L}_{\mathrm{amp}}\hat{\rho}. \tag{B18}$$

The first term again describes the measurement tone used for cavity readout, and the second term describes cavity losses. However, the cavity's open port is now directed to a phase-preserving amplifier downstream. The superoperator $\mathcal{L}_{\mathrm{amp}}$ is the Liouvillian defining this quantum amplifier, which we take to be a two-mode nondegenerate parametric amplifier providing phase-preserving gain:

$$\begin{aligned}
\mathcal{L}_{\mathrm{amp}}\hat{\rho}_c = -i\left[\frac{-ig_{\mathrm{amp}}}{2}\hat{d}\hat{c} + \mathrm{H.c.}, \hat{\rho}_c\right] \\
+ \gamma_d\mathcal{D}[\hat{d}]\hat{\rho}_c + \gamma\mathcal{D}[\hat{c}]\hat{\rho}_c.
\end{aligned} \tag{B19}$$

The superoperator $\mathcal{L}_c$ then defines the nonreciprocal coupling between the cavity mode and the amplifier's signal mode $\hat{d}$:

$$\mathcal{L}_c\hat{\rho}_c = -i\left[\frac{ig}{2}\hat{d}\hat{a}^\dagger + \mathrm{H.c.}, \hat{\rho}_c\right] + \Gamma\mathcal{D}[\hat{a} + \hat{d}]\hat{\rho}_c. \tag{B20}$$

To ensure nonreciprocal coupling so that fields from the cavity that carry qubit state information are transmitted to

the amplifier for readout, but transmission in the reverse direction is forbidden, we require $g = \Gamma$ [59].

Finally, $\mathcal{S}[dW]$ describes conditional evolution under continuous heterodyne monitoring, now of the amplifier's signal mode:

$$\mathcal{S}[dW]\hat{\rho}_c = \sqrt{\frac{\gamma_d}{2}}\left(\hat{d}\hat{\rho}_c + \hat{\rho}_c\hat{d}^\dagger - \langle\hat{d}+\hat{d}^\dagger\rangle\hat{\rho}_c\right)dW_I + \sqrt{\frac{\gamma_d}{2}}\left(-i\hat{d}\hat{\rho}_c + i\hat{\rho}_c\hat{d}^\dagger - \langle -i\hat{d}+i\hat{d}^\dagger\rangle\hat{\rho}_c\right)dW_Q. \tag{B21}$$

We now summarize the actual parameter choices used to generate quantum amplifier simulated datasets in the main text. We define the total cavity loss rate $\kappa = \kappa' + \Gamma$. Then we choose cavity parameters so that $\kappa' = \Gamma = 0.5\kappa$ and the dispersive shift $\chi/\kappa = 0.5$. Recall that perfect nonreciprocal coupling in the desired direction requires $g = \Gamma = 0.5\kappa$. Lastly, amplifier parameters are chosen so that $\gamma = \gamma_d + \Gamma = 5\kappa$, yielding the ratio of cold amplifier linewidth to cavity linewidth $\gamma/\kappa = 5$ used in the main text, and also implying that $\gamma_d = 4.5\kappa$.

In the absence of qubit transitions, Eq. (B8) holds once again, as $\mathcal{L}_{\text{meas}}$ is completely independent of the qubit sector. Hence, this sector may be traced out exactly as in the previous subsection. We thus arrive at a *cavity-amplifier-only* SME in the absence of any qubit transitions:

$$d\hat{\varrho}_c = -i[\hat{\mathcal{H}}_{\text{cav}}, \hat{\varrho}_c]dt + \mathcal{L}_{\text{meas}}[dW]\hat{\varrho}_c \tag{B22}$$

for $\mathcal{L}_{\text{meas}}$ now given by Eq. (B17). The resulting SME again preserves Gaussian states and can be solved exactly by a truncated equations of motion approach.

3. Dispersive readout including multilevel transitions using a cavity

For qubit readout allowing for state transitions, we must now include $\mathcal{L}_{\text{envt}}$ in the SME:

$$d\hat{\rho}_c = \mathcal{L}_{\text{sys}}\hat{\rho}_c\, dt + \mathcal{L}_{\text{envt}}\hat{\rho}_c + \mathcal{L}_{\text{meas}}[dW]\hat{\rho}_c. \tag{B23}$$

Again $\mathcal{L}_{\text{sys}}$ is given by Eq. (B2). The nontrivial superoperator $\mathcal{L}_{\text{envt}}$ takes the form

$$\mathcal{L}_{\text{envt}}\hat{\rho} = \sum_{j\neq k}\gamma_{jk}\mathcal{D}[|k\rangle\langle j|]\hat{\rho}, \tag{B24}$$

where γ_{jk} is the rate of transition from qubit state $|j\rangle$ to qubit state $|k\rangle$.

As we still consider readout using a cavity, the remaining terms in Eq. (B23) are as in Eq. (B25); in particular, $\mathcal{L}_{\text{meas}}$ takes the form

$$\mathcal{L}_{\text{meas}}[dW]\hat{\rho}_c = \mathcal{L}_{\text{q}}\hat{\rho}_c + \mathcal{S}[dW]\hat{\rho}_c, \tag{B25}$$

where $\mathcal{L}_{\text{q}}$ is given by

$$\mathcal{L}_{\text{q}}\hat{\rho} = -i[\eta(\hat{a}+\hat{a}^\dagger), \hat{\rho}] + \kappa\mathcal{D}[\hat{a}]\hat{\rho}, \tag{B26}$$

while $\mathcal{S}[dW]$ is given by

$$\mathcal{S}[dW]\hat{\rho}_c = \sqrt{\frac{\kappa}{2}}\left(\hat{a}\hat{\rho}_c + \hat{\rho}_c\hat{a}^\dagger - \langle\hat{a}+\hat{a}^\dagger\rangle\hat{\rho}_c\right)dW_I + \sqrt{\frac{\kappa}{2}}\left(-i\hat{a}\hat{\rho}_c + i\hat{\rho}_c\hat{a}^\dagger - \langle -i\hat{a}+i\hat{a}^\dagger\rangle\hat{\rho}_c\right)dW_Q. \tag{B27}$$

We emphasize that now the quantum state of the measurement chain cannot generally be expressed in the form of Eq. (B8). Hence, Eq. (B23) is integrated in the joint qubit-cavity Hilbert space to generate simulated measurement datasets.

APPENDIX C: TRAINING AND TESTING DETAILS

In this appendix, we analyze how optimal weights $\mathbf{W}^{\text{opt}}$ are learned from a training dataset in the TPP approach.

1. Cost function and learned weights

We begin with the TPP map defined in the main text, Eq. (1),

$$\sigma^{\text{est}} = F\left[\mathbf{y}_{(n)}\right] = F\left[\mathbf{W}\vec{\boldsymbol{x}}_{(n)} + \mathbf{b}\right], \tag{C1}$$

now written to describe the mapping of a single instance n of measured data, compiled in the vector $\vec{x}_{(n)}$, to a vector $\mathbf{y}_{(n)} \in \mathbb{R}^C$. The mapping is via a set of weights $\mathbf{W}$ applied linearly to the data $\vec{x}$ and a set of weights that are additive, compiled in a column vector of biases $\mathbf{b} \in \mathbb{R}^C$.

The vector $\vec{\boldsymbol{x}}$ lives in the *joint* space of measurement records: $\vec{x}_{(n)} \in \mathbb{R}^{N_{\text{O}}N_{\text{T}}}$ is also a column vector and can be written in the form

$$\vec{\boldsymbol{x}}_{(n)} = \begin{pmatrix} \vec{x}_{1(n)} \\ \vec{x}_{2(n)} \\ \vdots \\ \vec{x}_{N_{\text{O}}(n)} \end{pmatrix}, \tag{C2}$$

where each vector $\vec{x}_{m(n)} \in \mathbb{R}^{N_{\text{T}}}$ is a column vector describing the discretized records of $m \in [N_{\text{O}}]$ measurement observables, each with N_{T} samples. Recall that for standard heterodyne readout, $N_{\text{O}} = 2$, where $\vec{x}_1 = \vec{I}$ and $\vec{x}_2 = \vec{Q}$. From here on, we can work with this concatenated vector $\vec{\boldsymbol{x}}$.

In Eq. (C1), $F[\cdot]$ is a function that maps the vector of measured heterodyne records to a discrete, scalar state label $\sigma \in [1, \ldots, C]$. This mapping is done via two operations. First, the measurement records $\vec{\boldsymbol{x}}_{(n)}^{(\sigma)}$ are mapped to an intermediate target vector $\mathbf{y}_{(n)}^{(\sigma)}$ by means of a "one-hot" encoding (conventional for classification tasks). The k th element of this target vector $\mathbf{y}_{(n)}^{(\sigma)}$ is given by

$$[\mathbf{y}_{(n)}^{(\sigma)}]_k = \begin{cases} 1 & \text{if } k = \sigma, \\ 0 & \text{otherwise.} \end{cases} \tag{C3}$$

Finally, a discriminator is used to map the target vector to a scalar state label.

With the key notation in place, we can discuss how the TPP training dataset is constructed. A training dataset of size N_{train} consists of $n \in [N_{\text{train}}]$ heterodyne records for each of the C states required to be distinguished in the classification task. We define a matrix $\mathbf{X} \in \mathbb{R}^{N_{\text{O}}N_{\text{T}} \times CN_{\text{train}}}$:

$$\mathbf{X} = \left(\vec{\boldsymbol{x}}_{(1)}^{(1)} \quad \vec{\boldsymbol{x}}_{(2)}^{(1)} \quad \cdots \quad \vec{\boldsymbol{x}}_{(N_{\text{train}})}^{(1)} \quad \cdots \quad \vec{\boldsymbol{x}}_{(1)}^{(C)} \quad \vec{\boldsymbol{x}}_{(2)}^{(C)} \quad \cdots \quad \vec{\boldsymbol{x}}_{(N_{\text{train}})}^{(C)}\right). \tag{C4}$$

We also define a matrix $\mathbf{Y} \in \mathbb{R}^{C \times CN_{\text{train}}}$ compiling the corresponding targets:

$$\mathbf{Y} = \left(\mathbf{y}_{(1)}^{(1)} \quad \mathbf{y}_{(2)}^{(1)} \quad \cdots \quad \mathbf{y}_{(N_{\text{train}})}^{(1)} \quad \cdots \quad \mathbf{y}_{(1)}^{(C)} \quad \mathbf{y}_{(2)}^{(C)} \quad \cdots \quad \mathbf{y}_{(N_{\text{train}})}^{(C)}\right). \tag{C5}$$

By further introducing $\vec{1} \in \mathbb{R}^{1 \times CN_{\text{train}}}$ as a row vector containing all 1's, Eq. (C1) for all CN_{train} records per measured observable can be written in the compact matrix form

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}\vec{1}. \tag{C6}$$

Before proceeding, we note that we have the freedom to introduce any invertible matrix $\mathbf{L} \in \mathbb{R}^{N_{\text{O}}N_{\text{T}} \times N_{\text{O}}N_{\text{T}}}$ as follows, without modifying the TPP map:

$$\mathbf{Y} = \mathbf{W}(\mathbf{L}^{-1}\mathbf{L})\mathbf{X} + \mathbf{b}\vec{1} = \begin{pmatrix}\mathbf{W}\mathbf{L}^{-1} & \mathbf{b}\end{pmatrix}\begin{pmatrix}\mathbf{L}\mathbf{X} \\ \vec{1}\end{pmatrix} \equiv \boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{X}}. \tag{C7}$$

The auxiliary matrix $\mathbf{L}$ will prove convenient for our analysis later.

Equation (C7) helps us define $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{(N_{\text{O}}N_{\text{T}}+1) \times CN_{\text{train}}}$ as a matrix that contains all measured records as well as a row of 1's to account for the contribution of biases. Then $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{C \times (N_{\text{O}}N_{\text{T}}+1)}$ is the composite matrix of all learned weights. Equation (C7) defines a regression problem that can be solved to obtain the optimal weights [60],

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{Y}\boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T)^{-1}. \tag{C8}$$

For convenience of the analysis to follow, we introduce two new matrices: the *mean* matrix $\mathbf{M} \in \mathbb{R}^{C \times (N_{\text{O}}N_{\text{T}}+1)}$,

$$N_{\text{train}}\mathbf{M} \equiv \mathbf{Y}\boldsymbol{\mathcal{X}}^T, \tag{C9}$$

and the *second-order moments matrix* $\mathbf{C} \in \mathbb{R}^{(N_{\text{O}}N_{\text{T}}+1) \times (N_{\text{O}}N_{\text{T}}+1)}$,

$$N_{\text{train}}\mathbf{C} \equiv \boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T, \tag{C10}$$

so that Eq. (C8) can equivalently be written as

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{M}\mathbf{C}^{-1}, \tag{C11}$$

where the factors of N_{train} cancel out.

Note that the matrix $\mathbf{C} = \boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T$ can at times be ill-conditioned, making its inverse difficult to compute numerically. In such cases, we instead compute the quantity $\mathbf{C}^+$, which is related to the pseudoinverse of $\boldsymbol{\mathcal{X}}$ and is given by the following limit relation defining the pseudoinverse:

$$\mathbf{C}^+ = \lim_{\lambda \to 0}(\mathbf{C} - \lambda\mathbf{I})^{-1}, \tag{C12}$$

where $\mathbf{I}$ is the identity matrix on $\mathbb{R}^{(N_{\text{O}}N_{\text{T}}+1) \times (N_{\text{O}}N_{\text{T}}+1)}$ and λ is typically referred to as a regularization parameter. If $\mathbf{C}$ is invertible, we have $\mathbf{C}^+ \to \mathbf{C}^{-1}$. We emphasize that for the datasets analyzed in this paper, the intrinsic dataset noise serves as an effective regularizer, such that we can typically set $\lambda = 0$.

2. Testing via cross-validation

For all classification infidelities calculated in the main text, we perform cross-validation. For a full dataset of N_{traj} records per state, a training set is constructed with $N_{\text{train}} < N_{\text{traj}}$ records as described above. The remaining $N_{\text{test}} =$

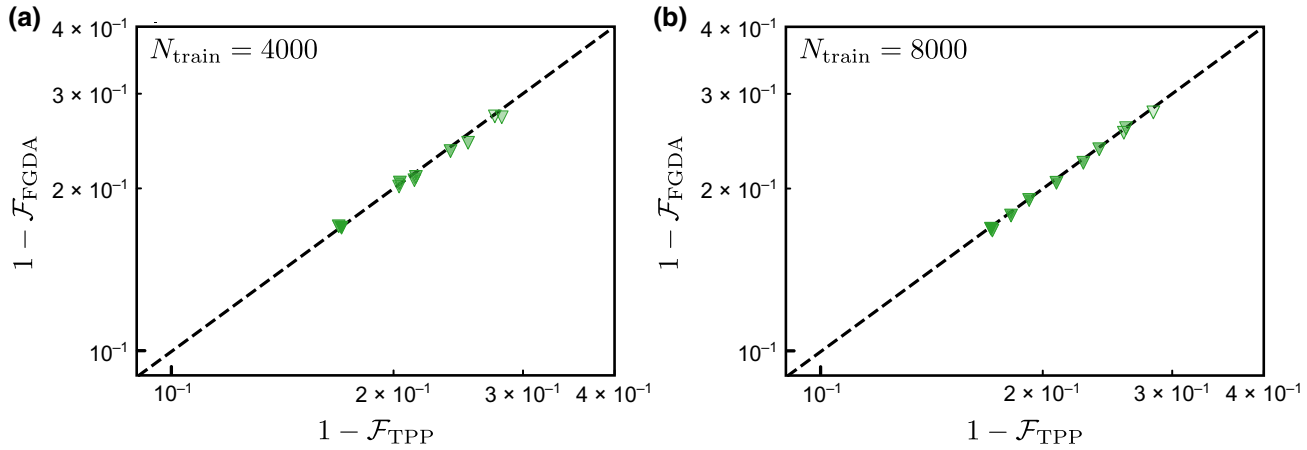


FIG. 9. Comparison of general TPP versus FGDA as a function of training set size and qubit initialization error. Results are shown for training set sizes N_{train} of (a) 4000 measurement records per class and (b) 8000 measurement records per class. More-opaque markers indicate lower qubit initialization error.

$N_{\text{traj}} - N_{\text{train}}$ records are used to construct a testing set. We use 80% of the dataset for training and the remaining 20% for testing. This is consistent with training set and testing set sizes for standard machine-learning applications; for example, the MNIST handwritten digits classification task [61] uses 85.7% of the total dataset for training and the remaining 14.3% for testing. Predicted state labels are obtained with this testing set via both the FGDA scheme, Eq. (10), and the TPP, Eq. (1). The process is repeated until a total of $L = 10$ iterations are completed: each time, a new set of weights $\mathcal{W}^{\text{opt}}$ is obtained from a distinct randomly chosen training set of the total N_{traj} records, and classification infidelities are computed with the new random testing datasets. All classification fidelities are averaged to obtain the final values plotted in the main text. This cross-validation approach is standard in machine learning and ensures that the observed performance is not unduly affected by variations due to the specific training dataset or testing dataset used.

3. Dependence on size and fidelity of training sets

As the TPP deploys a supervised learning approach to training [not unlike a standard matched filter, as shown by Eq. (11)], an important question is how its performance depends on the size of the available training dataset, as well as any possible errors in the labeling of data such as may arise due to qubit initialization errors for the case of qubit state readout.

To answer these questions, we consider again the case of measurement data experiencing only Gaussian white noise, and compare the general TPP performance and the FGDA performance for binary classification of $p \in \{e, g\}$. Note that use of only the *general* TPP makes sense here, as the white noise TPP is exactly equal to the standard matched filter learned from a given training dataset in the special case of white noise. In Fig. 9(a), we thus plot the performance of the general TPP against the performance of the FGDA as in the main text, with a training set size N_{train} of 4000 measurement records per class. We also consider the impact of qubit ground state $|g\rangle$ initialization error from 5% up to 35%: more-opaque markers correspond to a lower qubit initialization error and hence better classification performance. Figure 9(b) shows the same plot but now for a larger training set with $N_{\text{train}} = 8000$ measurement records per class.

For the smaller training dataset, the TPP very marginally underperforms in comparison with the FGDA for some of the data points. This is because the TPP has not yet converged to the optimal filter for this size of training dataset. With increasing training set size, this difference becomes smaller and smaller. We also note that qubit initialization error appears to impact both schemes similarly, so the TPP appears to not be unduly impacted by data mislabeling.

Finally, we emphasize that the task considered here is one that most heavily favors the standard FGDA in contrast to the general TPP, as the measurement data actually satisfy the noise conditions assumed *a priori* by the standard matched filter. Furthermore, the signal amplitudes used are weak (as indicated by the relatively low classification fidelity), so the measured data have a low SNR and more data are needed to probe the statistics faithfully. If the true measured data exhibit noise statistics that deviate from this white noise case, even TPP filters learned with use of small training set sizes can already outperform the then suboptimal standard MF trained on the same dataset.

APPENDIX D: TPP LEARNED WEIGHTS AS OPTIMAL FILTERS: ANALYTIC RESULTS

In this appendix, we attempt to find an explicit form for the matrix $\mathcal{W}^{\text{opt}}$ from Appendix C, under some assumptions on the form of the data contained in $\mathbf{X}$.

1. Measured data as stochastic random variables

To make further progress, we must make some assumptions regarding the general form of measured data $\vec{\boldsymbol{x}}^{(c)}$. In particular, we assume that

$$\vec{\boldsymbol{x}}^{(c)} = \vec{\boldsymbol{s}}^{(c)} + \vec{\boldsymbol{\xi}}^{(c)}, \tag{D1}$$

where $\vec{\boldsymbol{\xi}}^{(c)}$ is a random noise process that contains the stochasticity of the data $\vec{\boldsymbol{x}}$. In particular, this includes contributions from heterodyne measurement noise $\vec{\xi}$, added classical noise $\vec{\xi}_{\rm cl}$, and quantum noise in conditional quantum trajectories. Without loss of generality, $\vec{\boldsymbol{\xi}}$ can always be taken to have zero mean,

$$\mathbb{E}[\vec{\boldsymbol{\xi}}_j] = 0 \text{ for all } j\,. \tag{D2}$$

The random noise process can be defined by its covariance matrix,

$$\boldsymbol{\Sigma}^{(c)}_{jk} = \mathbb{E}[\vec{\boldsymbol{\xi}}^{(c)}_j \vec{\boldsymbol{\xi}}^{(c)}_k]. \tag{D3}$$

The noise process will, in general, also possess nonzero higher-order cumulants, but these quantities will not make an appearance in our analysis here.

Then, $\vec{\boldsymbol{s}}^{(c)}$ is simply equal to the expectation value of the random variable $\vec{\boldsymbol{x}}^{(c)}$ over an in principle infinite number of shots,

$$\vec{\boldsymbol{s}}^{(\boldsymbol{c})} = \mathbb{E}[\vec{\boldsymbol{x}}^{(c)}]. \tag{D4}$$

In practice, we will have access to only a finite number of shots $N_{\rm train}$. Then, the above mean can be approximated with use of the estimator

$$\vec{\boldsymbol{s}}^{(c)} \approx \frac{1}{N_{\rm train}} \sum_{n=1}^{N_{\rm train}} \vec{\boldsymbol{x}}^{(c)}_{(n)}. \tag{D5}$$

Similarly, the covariance matrix of the noise process can be estimated via

$$\boldsymbol{\Sigma}^{(c)} \approx \frac{1}{N_{\rm train}} \sum_{n=1}^{N_{\rm train}} \vec{\boldsymbol{\xi}}^{(c)}_{(n)} \vec{\boldsymbol{\xi}}^{(c)T}_{(n)}. \tag{D6}$$

Assuming we have the very general form of Eq. (D1), we can proceed to greatly simplify the matrices **M** and **C**.

a. Simplification of mean matrix M

The mean matrix **M**, Eq. (C9), can be written explicitly as

$$N_{\rm train}\mathbf{M} = \mathbf{Y}\begin{pmatrix}\mathbf{X}^T\mathbf{L}^T & \vec{1}^T\end{pmatrix} = \begin{pmatrix}\mathbf{Y}\mathbf{X}^T\mathbf{L}^T & \mathbf{Y}\vec{1}^T\end{pmatrix}. \tag{D7}$$

We now proceed to simplify the general matrices $\mathbf{Y}\vec{1}^T$ and $\mathbf{Y}\mathbf{X}^T\mathbf{L}^T$. Starting with the former, which simply yields a column vector that is an element of $\mathbb{R}^{C\times 1}$, we find explicitly

$$\begin{aligned}(\mathbf{Y}\vec{1}^T)_l &= \sum_{k=1}^{C\cdot N_{\rm train}} \mathbf{Y}_{lk}\vec{1}^T_k = \sum_n \sum_c \mathbf{y}^{(c)}_l \\ &= \sum_n \sum_c \delta_{cl} = N_{\rm train}.\end{aligned} \tag{D8}$$

Here we have used the fact that the sum over the columns of **Y** (and of **X**), indexed by k, can be decomposed into two sums: over $N_{\rm train}$ training records indexed by n and over C states indexed by c. From here on, we suppress the limits of these summations for clarity.

Next we consider $\mathbf{Y}\mathbf{X}^T$, which can be expanded out explicitly,

$$\begin{aligned}\mathbf{Y}\mathbf{X}^T &= \sum_k \mathbf{Y}_{lk}\mathbf{X}^T_{km} = \sum_k \mathbf{Y}_{lk}\mathbf{X}_{mk} \\ &= \sum_{n=1}^{N_{\rm train}} \sum_{c=1}^{C} \delta_{lc}(\vec{\boldsymbol{x}}^{(c)}_{(n)})_m \simeq N_{\rm train} \sum_{c=1}^{C} \delta_{lc}(\vec{\boldsymbol{s}}^{(c)})_m \\ &= N_{\rm train}(\vec{\boldsymbol{s}}^{(l)})_m,\end{aligned} \tag{D9}$$

where we have used Eq. (D5) in obtaining the final expression. Hence, with use of Eq. (D7), the matrix **M** takes the simple form (after the factors of $N_{\rm train}$ cancel out)

$$\mathbf{M} = \begin{pmatrix}(\mathbf{L}\vec{\boldsymbol{s}}^{(1)})^T & 1 \\ \vdots & \vdots \\ (\mathbf{L}\vec{\boldsymbol{s}}^{(C)})^T & 1\end{pmatrix} \equiv \begin{pmatrix}(\vec{S}^{(1)})^T \\ \vdots \\ (\vec{S}^{(C)})^T\end{pmatrix}, \tag{D10}$$

which contains the mean traces for all measured observables over all states, explaining the nomenclature of the mean matrix. We have further introduced the vectors $\vec{S}^{(c)}$, which also include the contribution from the bias.

b. Simplification of second-order moments matrix C

Simplifying the second-order correlation matrix **C** is more involved. We begin by expanding it to the form

$$\begin{aligned}N_{\rm train}\mathbf{C} \equiv \boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T &= \begin{pmatrix}\mathbf{L}\mathbf{X} \\ \vec{1}\end{pmatrix}\begin{pmatrix}\mathbf{X}^T\mathbf{L}^T & \vec{1}^T\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{L}\mathbf{X}\mathbf{X}^T\mathbf{L}^T & \mathbf{L}\mathbf{X}\vec{1}^T \\ \vec{1}\mathbf{X}^T\mathbf{L}^T & \vec{1}\vec{1}^T\end{pmatrix}.\end{aligned} \tag{D11}$$

Note that $\mathbf{X}\mathbf{X}^T$ is simply the two-time correlation matrix of the measured data.

We can further simplify $\mathbf{C}$, which has four components. Starting with the simplest, we note that

$$\vec{1}\vec{1}^T = \sum_k \vec{1}_k \vec{1}_k^T = \sum_n \sum_c 1 = CN_{\text{train}}. \tag{D12}$$

Next we consider the off-diagonal block term,

$$\begin{aligned}(\mathbf{X}\vec{1}^T)_i &= \sum_k (\mathbf{X})_{ik}(\vec{1}^T)_k = \sum_c \sum_n (\vec{\boldsymbol{x}}_{(n)}^{(c)})_i \\ &\simeq N_{\text{train}} \sum_c [\vec{\boldsymbol{s}}^{m(c)}]_i. \end{aligned} \tag{D13}$$

The other off-diagonal term is simply the transpose of the above.

Finally, we consider the block matrix,

$$\begin{aligned}[\mathbf{X}\mathbf{X}^T]_{ij} &= \sum_k [\mathbf{X}]_{ik}[\mathbf{X}^T]_{kj} = \sum_k [\mathbf{X}]_{ik}[\mathbf{X}]_{jk} \\ &= \sum_c \sum_n [\vec{\boldsymbol{x}}_{(n)}^{(c)}]_i [\vec{\boldsymbol{x}}_{(n)}^{(c)}]_j . \end{aligned} \tag{D14}$$

To proceed further, we substitute Eq. (D1) into the final expression and expand it:

$$\begin{aligned}[\mathbf{X}\mathbf{X}^T]_{ij} &= \sum_c \sum_n [\vec{\boldsymbol{x}}_{(n)}^{(c)}]_i [\vec{\boldsymbol{x}}_{(n)}^{(c)}]_j \\ &= \sum_c \Bigg\{ [\vec{\boldsymbol{s}}^{(c)}]_i [\vec{\boldsymbol{s}}^{(c)}]_j + \sum_n [\vec{\boldsymbol{\xi}}_{(n)}^{(c)}]_i [\vec{\boldsymbol{s}}^{(c)}]_j \\ &\quad + [\vec{\boldsymbol{s}}^{(c)}]_i \sum_n [\vec{\boldsymbol{\xi}}_{(n)}^{I(c)}]_j + \sum_n [\vec{\boldsymbol{\xi}}_{(n)}^{(c)}]_i [\vec{\boldsymbol{\xi}}_{(n)}^{(c)}]_j . \Bigg\} \end{aligned} \tag{D15}$$

Note that the sums indexed by n over the training data are estimators of the statistics of the noise process. We can therefore write

$$[\mathbf{X}\mathbf{X}^T]_{ij} = N_{\text{train}} \sum_c \left\{ [\vec{\boldsymbol{s}}^{(c)}]_i [\vec{\boldsymbol{s}}^{(c)}]_j + \boldsymbol{\Sigma}_{ij}^{(c)} \right\}. \tag{D16}$$

It now proves useful to introduce two further matrices, the *Gram* matrix $\mathbf{G}$,

$$\mathbf{G} = \sum_c \vec{\boldsymbol{s}}^{(c)} (\vec{\boldsymbol{s}}^{(c)})^T, \tag{D17}$$

and the empirical *correlation* matrix $\mathbf{V}$,

$$\mathbf{V} = \sum_c \boldsymbol{\Sigma}^{(c)}. \tag{D18}$$

We can therefore write $\mathbf{C}$ in the simplified form

$$\mathbf{C} = \begin{pmatrix} \mathbf{L}(\mathbf{G}+\mathbf{V})\mathbf{L}^T & \sum_c \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ \sum_c (\vec{\boldsymbol{s}}^{(c)})^T \mathbf{L}^T & C \end{pmatrix} \tag{D19}$$

and hence construct the full $\mathbf{C}$ via Eq. (D11).

Having constructed explicit forms of $\mathbf{M}$ and $\mathbf{C}$, we are, in principle, positioned to evaluate the optimal weights and biases $\mathcal{W}^{\text{opt}}$ explicitly as well. To do so, it first again proves useful to interpret the learned weights in terms of optimal filters.

2. Constraints on TPP filters

The learned matrix of weights can be written in vector form as

$$\mathcal{W}^{\text{opt}} \equiv \begin{pmatrix} (\vec{\boldsymbol{f}}_1)^T \mathbf{L}^{-1} & b_1 \\ \vdots & \vdots \\ (\vec{\boldsymbol{f}}_C)^T \mathbf{L}^{-1} & b_C \end{pmatrix} \equiv \begin{pmatrix} (\vec{F}_1)^T \\ \vdots \\ (\vec{F}_C)^T \end{pmatrix}. \tag{D20}$$

Next, using Eq. (C8) together with the explicit form of the mean matrix $\mathbf{M}$ in Eq. (D10), we arrive at the important relation

$$\begin{aligned}&\begin{pmatrix} (\vec{F}_1)^T \\ \vdots \\ (\vec{F}_C)^T \end{pmatrix} = \begin{pmatrix} (\vec{S}^{(1)})^T \\ \vdots \\ (\vec{S}^{(C)})^T \end{pmatrix} \mathbf{C}^{-1} \implies \\ &\mathbf{C}^{-1} \left(\vec{S}^{(1)} \cdots \vec{S}^{(C)} \right) = \left(\vec{F}_1 \cdots \vec{F}_C \right), \end{aligned} \tag{D21}$$

where we have used the fact that $\mathbf{C}$, and hence its inverse, is a symmetric matrix, and thereby computed the transpose of both sides. The above equation then implies

$$\mathbf{C}^{-1} \vec{S}^{(c)} = \vec{F}_c, \tag{D22}$$

We note that the matrix $\mathbf{C}$ is very general as it is constructed for completely arbitrary measured signals; it is therefore generally dense and its inverse $\mathbf{C}^{-1}$ cannot be analytically determined. However, Eq. (D22) suggests that if we can find a way to work with quantities $\mathbf{C}^{-1}\vec{S}^{(c)}$ directly, we can avoid having to evaluate this regularized inverse of $\mathbf{C}$. This is our strategy to evaluate optimal filters analytically.

We demonstrate this approach by considering the action of $\mathbf{C}$ on the constant inhomogeneous vector,

$$\vec{n} = \begin{pmatrix} \vec{\mathbf{0}} \\ 1 \end{pmatrix}, \tag{D23}$$

where $\vec{\mathbf{0}} \in \mathbb{R}^{N_O N_T}$ is a vector of 0's. In particular, we wish to evaluate $\mathbf{C}\vec{n}$. Using the block representation of $\mathbf{C}$, we

have

$$\mathbf{C}\vec{n} = \begin{pmatrix} \mathbf{L}(\mathbf{G}+\mathbf{V})\mathbf{L}^T & \sum_c \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ \sum_c (\vec{\boldsymbol{s}}^{(c)})^T \mathbf{L}^T & C \end{pmatrix} \begin{pmatrix} \vec{\mathbf{0}} \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_c \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ C \end{pmatrix}$$
$$= \sum_c \begin{pmatrix} \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ 1 \end{pmatrix} = \sum_c \vec{S}^{(c)}. \tag{D24}$$

Most importantly, note that the right-hand side is entirely independent of the covariance matrix $\mathbf{V}$, instead depending only on mean traces.

Multiplying Eq. (D24) by $\mathbf{C}^{-1}$ and then making use of Eq. (D22) will allow us to work directly with the (unknown) optimal filters $\vec{F}_c$. We immediately find

$$\sum_c \vec{F}_c = \vec{n}. \tag{D25}$$

For completeness, we also consider the case where we instead require the calculation of $\mathbf{C}^{+}$. To this end, we add and subtract the regularization parameter λ,

$$(\mathbf{C}-\lambda\mathbf{I})\vec{n} + \lambda\vec{n} = \sum_c \vec{S}^{(c)} \implies \sum_c (\mathbf{C}-\lambda)^{-1}\vec{S}^{(c)}$$
$$= \vec{n} + \lambda(\mathbf{C}-\lambda\mathbf{I})^{-1}\vec{n}, \tag{D26}$$

or, finally,

$$\sum_c \vec{F}_c = \vec{n} + \lambda(\mathbf{C}-\lambda\mathbf{I})^{-1}\vec{n}. \tag{D27}$$

The above defines a constraint on learned optimal filters, implying that they are not all linearly independent. Crucially, this constraint holds regardless of the correlation properties of the noise characterized by $\mathbf{V}$ and is hence very general.

3. Analytically calculable TPP filters: "Matched filters" for arbitrary *C*

Having obtained a useful constraint on TPP-learned filters, we now take a step further and calculate semianalytic expressions for these learned filters [eventually arriving at Eq. (14)].

The first step is to simplify the form of the matrix $\mathbf{C}$ in Eq. (D19), which we reproduce and expand below:

$$\mathbf{C} = \begin{pmatrix} \mathbf{L}\mathbf{G}\mathbf{L}^T + \mathbf{L}\mathbf{V}\mathbf{L}^T & \sum_c \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ \sum_c (\vec{\boldsymbol{s}}^{(c)})^T \mathbf{L}^T & C \end{pmatrix}. \tag{D28}$$

We have thus far allowed the auxiliary matrix $\mathbf{L}$ to be completely general; we can now use it to simplify the form of $\mathbf{C}$. Note that $\mathbf{V}$ as defined in Eq. (D18) is the positive sum of individual positive-definite correlation matrices; as a result, it must also be positive-definite and real. Among the useful properties of such positive-definite matrices is that they admit a Cholesky decomposition. We choose the auxiliary matrix $\mathbf{L}$ such that it precisely determines the Cholesky decomposition of $\mathbf{V}$:

$$\mathbf{V} = \mathbf{L}^{-1}(\mathbf{L}^T)^{-1} \implies \mathbf{V}^{-1} = \mathbf{L}^T\mathbf{L}, \tag{D29}$$

where we have also used the fact that a positive-definite matrix is always invertible.

With this choice, we immediately find that $\mathbf{C}$ reduces to

$$\mathbf{C} = \begin{pmatrix} \mathbf{L}\mathbf{G}\mathbf{L}^T + \bar{\mathbf{I}} & \sum_c \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ \sum_c (\vec{\boldsymbol{s}}^{(c)})^T \mathbf{L}^T & C \end{pmatrix}, \tag{D30}$$

where $\bar{\mathbf{I}}$ is the identity matrix on $\mathbb{R}^{N_O N_T \times N_O N_T}$.

a. Obtaining the linear system for filters

To obtain a system of equations for the learned filters, we now consider the action of $\mathbf{C}$ on the vector $\vec{S}^{(c)}$. To do so, we once again make use of the simplified block representation of $\mathbf{C}$, which allows us to write

$$\mathbf{C}\vec{S}^{(c)} = \begin{pmatrix} \sum_{c'} \mathbf{L}\vec{\boldsymbol{s}}^{(c')}(\vec{\boldsymbol{s}}^{(c')})^T\mathbf{L}^T + \bar{\mathbf{I}} & \sum_{c'} \mathbf{L}\vec{\boldsymbol{s}}^{(c')} \\ \sum_{c'} (\vec{\boldsymbol{s}}^{(c')})^T\mathbf{L}^T & C \end{pmatrix} \begin{pmatrix} \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ 1 \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{c'} \mathbf{L}\vec{\boldsymbol{s}}^{(c')}[(\vec{\boldsymbol{s}}^{(c')})^T\mathbf{L}^T\mathbf{L}\vec{\boldsymbol{s}}^{(c)}] + \mathbf{L}\vec{\boldsymbol{s}}^{(c)} + \sum_{c'} \mathbf{L}\vec{\boldsymbol{s}}^{(c')} \\ \sum_{c'} [(\vec{\boldsymbol{s}}^{(c')})^T\mathbf{L}^T\mathbf{L}\vec{\boldsymbol{s}}^{(c)}] + C \end{pmatrix}. \tag{D31}$$

It proves useful to define the overlap of mean traces,

$$O_{cc'} = (\vec{\boldsymbol{s}}^{(c')})^T\mathbf{L}^T\mathbf{L}\vec{\boldsymbol{s}}^{(c)} = (\vec{\boldsymbol{s}}^{(c')})^T\mathbf{V}^{-1}\vec{\boldsymbol{s}}^{(c)}, \tag{D32}$$

where we have used Eq. (D29). We can thus write

$$\mathbf{C}\vec{S}^{(c)} = \begin{pmatrix} \sum_{c'} O_{cc'}\mathbf{L}\vec{\boldsymbol{s}}^{(c')} + \sum_{c'} \mathbf{L}\vec{\boldsymbol{s}}^{(c')} + \mathbf{L}\vec{\boldsymbol{s}}^{(c)} \\ \sum_{c'} O_{cc'} + C \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{c'} [O_{cc'} + 1 + \delta_{cc'}]\mathbf{L}\vec{\boldsymbol{s}}^{(c')} \\ \sum_{c'} [O_{cc'} + 1] \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{c'} [O_{cc'} + 1 + \delta_{cc'}]\mathbf{L}\vec{\boldsymbol{s}}^{(c')} \\ \sum_{c'} [O_{cc'} + 1 + \delta_{cc'}] \end{pmatrix} - \begin{pmatrix} \vec{\mathbf{0}} \\ \sum_{c'} \delta_{cc'} \end{pmatrix}$$
$$= \sum_{c'} [O_{cc'} + 1 + \delta_{cc'}] \begin{pmatrix} \mathbf{L}\vec{\boldsymbol{s}}^{(c')} \\ 1 \end{pmatrix} - \begin{pmatrix} \vec{\mathbf{0}} \\ 1 \end{pmatrix}. \tag{D33}$$

Finally, defining

$$M_{cc'} = [O_{cc'} + 1 + \delta_{cc'}] \tag{D34}$$

and once again introducing $\vec{n}$ from Eq. (D23), we arrive at the form

$$\mathbf{C}\vec{S}^{(c)} = \sum_{c'} M_{cc'}\vec{S}^{(c')} - \vec{n}. \tag{D35}$$

Therefore, we find that the action of $\mathbf{C}$ on $\vec{S}^{(c)}$ can be expressed as a linear combination of the set of vectors $\{\vec{S}^{(c)}\}$ and a vector $\vec{n}$ that is independent of c.

We now wish to introduce the unknown filters $\vec{F}_c$ to the above system using Eq. (D22). To do so, we add and subtract the regularization parameter λ, followed by multiplication by the regularized inverse of $\mathbf{C}$. This yields

$$\begin{aligned}\vec{S}^{(c)} &= \sum_{c'}(\mathbf{C}-\lambda\mathbf{I})^{-1}(M_{cc'}-\lambda\mathbf{I}\delta_{cc'})\vec{S}^{(c')} - (\mathbf{C}-\lambda\mathbf{I})^{-1}\vec{n} \\ &= \sum_{c'}(M_{cc'}-\lambda\mathbf{I}\delta_{cc'})\vec{F}_{c'} - (\mathbf{C}-\lambda\mathbf{I})^{-1}\vec{n}. \end{aligned} \tag{D36}$$

However, Eq. (D36) is not entirely free of the $(\mathbf{C}-\lambda\mathbf{I})^{-1}$ matrix, due to the inhomogeneous term. Fortunately, as the inhomogeneous term is constant, it can be removed by our considering the difference of Eq. (D36) for any two distinct c values. For example, for $c \neq c'' \in [1,\ldots,C]$,

$$\begin{aligned}\vec{S}^{(c)} - \vec{S}^{(c'')} &= \sum_{c'} M_{cc'}\vec{F}_{c'} - \sum_{c'} M_{c''c'}\vec{F}_{c'} \\ &= \sum_{c'}[M_{cc'} - M_{c''c'}]\vec{F}_{c'}. \end{aligned} \tag{D37}$$

This naturally introduces the difference of mean traces to the calculation of learned filters.

Finally, we recall that the unknown filters $\vec{F}_c$ are not all linearly independent. We therefore use the constraint expressed in Eq. (D25) in the formal limit $\lambda \to 0$ to eliminate one of the unknown vectors, here taken to be $\vec{F}_C$:

$$\vec{F}_C = \vec{n} - \sum_{c'=1}^{C-1}\vec{F}_{c'}. \tag{D38}$$

Then Eq. (D37) can be rewritten as

$$\begin{aligned}\vec{S}^{(c)} - \vec{S}^{(c'')} &= \sum_{c'=1}^{C-1}[M_{cc'} - M_{c''c'}]\vec{F}_{c'} + [M_{cC} - M_{c''C}]\vec{F}_C \\ &= \sum_{c'=1}^{C-1}[M_{cc'} - M_{c''c'}]\vec{F}_{c'} - \sum_{c'=1}^{C-1}[M_{cC} - M_{c''C}]\vec{F}_{c'} + [M_{cC} - M_{c''C}]\vec{n} \\ &= \sum_{c'=1}^{C-1}[(M_{cc'} - M_{c''c'}) - (M_{cC} - M_{c''C})]\vec{F}_{c'} + [M_{cC} - M_{c''C}]\vec{n}. \end{aligned} \tag{D39}$$

Note that there are $C-1$ unknowns $\vec{F}_c$, and hence we require $C-1$ equations. These equations are simply provided by Eq. (D39) by our considering $C-1$ distinct pairs $[c,c'']$. For concreteness, we consider pairs $P_p = [c,c'']$, where $[c,c''] \in \{[1,2],[2,3],\ldots,[C-1,C]\}$ indexed by $p \in [1,\ldots,C-1]$. We also introduce notation to individually identify the states constituting the p th pair, for convenience: if $P_p = [c,c'']$, $P_p(1) = c, P_p(2) = c''$. We then define the difference of mean traces constituting a pair,

$$\vec{S}^{P_p} \equiv \vec{S}^{(P_p(1))} - \vec{S}^{(P_p(2))}. \tag{D40}$$

Each pair yields an equation of the form of Eq. (D39); it is easily seen that the full set of $C-1$ equations can be compiled into the matrix system

$$\begin{pmatrix}\vec{S}^{P_1}\\ \vdots \\ \vec{S}^{P_{C-1}}\end{pmatrix} = (\mathbf{Q}\otimes\mathbf{I})\begin{pmatrix}\vec{F}_1\\ \vdots \\ \vec{F}_{C-1}\end{pmatrix} + (\mathbf{T}\otimes\mathbf{I})\begin{pmatrix}\vec{n}\\ \vdots \\ \vec{n}\end{pmatrix} \tag{D41}$$

with use of the properties of the Kronecker product. Here $\mathbf{I}$ is the identity matrix on $\mathbb{R}^{N_O(N_T+1)\times N_O(N_T+1)}$ as before, while $\mathbf{Q}$ and $\mathbf{T}$ are both elements of the much smaller space $\mathbb{R}^{(C-1)\times(C-1)}$. In particular, their matrix elements are given by

$$\mathbf{Q}_{pc} = \left[(M_{P_p(1)c} - M_{P_p(2)c}) - (M_{P_p(1)C} - M_{P_p(2)C})\right], \quad \mathbf{T}_{pc} = \delta_{pc}\left[M_{P_p(1)C} - M_{P_p(2)C}\right]. \tag{D42}$$

Note further that $\mathbf{T}$ is a diagonal matrix.

b. Solving the linear system for filters

Being a simple linear system, Eq. (D41) has the formal solution

$$\begin{pmatrix} \vec{F}_1 \\ \vdots \\ \vec{F}_{C-1} \end{pmatrix} = \left(\mathbf{Q}^{-1} \otimes \mathbf{I}\right) \begin{pmatrix} \vec{S}^{P_1} \\ \vdots \\ \vec{S}^{P_{C-1}} \end{pmatrix} - \left(\mathbf{Q}^{-1} \otimes \mathbf{I}\right)(\mathbf{T} \otimes \mathbf{I}) \begin{pmatrix} \vec{n} \\ \vdots \\ \vec{n} \end{pmatrix}. \tag{D43}$$

We can now simply read off the solution for the unknown vector $\vec{F}_c$:

$$\vec{F}_c = \sum_{p=1}^{C-1} \mathbf{Q}^{-1}_{cp} \vec{S}^{P_p} - \sum_{p=1}^{C-1} \mathbf{Q}^{-1}_{cp} \mathbf{T}_{pp} \vec{n}. \tag{D44}$$

The first term on the right-hand side completely defines the filter components in $\vec{F}_c$, as they have a zero at the position corresponding to the bias component. The second term then entirely defines the bias. Using the form of $\vec{F}_c$ from Eq. (D20), we can immediately read off the individual filters for each measured observable:

$$(\mathbf{L}^{-1})^T \vec{\boldsymbol{f}}_c = \sum_p \mathbf{Q}^{-1}_{cp} \mathbf{L} \vec{\boldsymbol{s}}^{(P_p)}, \tag{D45}$$

which simplifies to

$$\vec{\boldsymbol{f}}_c = \sum_p \mathbf{Q}^{-1}_{cp} \mathbf{L}^T \mathbf{L} \vec{\boldsymbol{s}}^{(P_p)} \implies \vec{\boldsymbol{f}}_c = \sum_p \mathbf{Q}^{-1}_{cp} \mathbf{V}^{-1} \vec{\boldsymbol{s}}^{(P_p)}, \tag{D46}$$

where we have again used Eq. (D29). The bias terms are finally given by

$$\mathbf{b}_c = -\sum_p \mathbf{Q}^{-1}_{cp} \mathbf{T}_{pp}. \tag{D47}$$

The remaining learned filter and bias are then given by the constraint, Eq. (D25).

An alternative, more practical form of the learned filters can be extracted by transitioning from the representation in terms of difference vectors $\vec{S}^{P_p}$ to the individual traces $\vec{S}^{(c)}$ with use of Eq. (D40). We find

$$\vec{\boldsymbol{f}}_c = \mathbf{Q}^{-1}_{c1} \mathbf{V}^{-1} \vec{\boldsymbol{s}}^{(1)} + \sum_{p=2}^{C-1} \left[\mathbf{Q}^{-1}_{cp} - \mathbf{Q}^{-1}_{c(p-1)}\right] \mathbf{V}^{-1} \vec{\boldsymbol{s}}^{(p)} - \mathbf{Q}^{-1}_{c(C-1)} \mathbf{V}^{-1} \vec{\boldsymbol{s}}^{(C)}, \tag{D48}$$

which provides the learned filters as a linear combination of mean signals corresponding to each state to be classified. From comparison with Eq. (14), we finally have

$$\vec{\boldsymbol{f}}_c = \sum_{p=1}^{C} C_{cp} \mathbf{V}^{-1} \vec{\boldsymbol{s}}^{(p)}, \quad C_{cp} = \begin{cases} +\mathbf{Q}^{-1}_{c1} & \text{if } p = 1, \\ -\mathbf{Q}^{-1}_{c(C-1)} & \text{if } p = C, \\ \mathbf{Q}^{-1}_{cp} - \mathbf{Q}^{-1}_{c(p-1)} & \text{otherwise.} \end{cases} \tag{D49}$$

4. Reduction to standard matched filter for binary classification ($C = 2$)

For $C = 2$, the matrix system in Eq. (D41) reduces to a single equation:

$$\vec{S}^{(1)} - \vec{S}^{(2)} = [M_{11} - M_{21} - (M_{12} - M_{22})]\,\vec{F}_1 + [M_{21} - M_{22}]\,\vec{n}. \tag{D50}$$

From here we can directly read off the filter and bias term:

$$\begin{pmatrix} \vec{f}_1 \\ \mathbf{b}_1 \end{pmatrix} = \frac{\mathbf{V}^{-1}}{M_{11} - M_{21} - (M_{12} - M_{22})} \begin{pmatrix} \vec{s}^{(1)} - \vec{s}^{(2)} \\ 0 \end{pmatrix} - \frac{M_{21} - M_{22}}{M_{11} - M_{21} - (M_{12} - M_{22})} \begin{pmatrix} \vec{0} \\ 1 \end{pmatrix}. \tag{D51}$$

5. Example: Analytic construction of TPP-learned filters for three-state classification ($C = 3$)

We now provide an example of the construction of TPP-learned optimal filters for $C = 3$ state classification. To compute these filters using Eq. (D49), we simply require knowledge of the matrix $\mathbf{Q}$, whose matrix elements are given by Eq. (D42). For $C = 3$, $\mathbf{Q} \in \mathbb{R}^{2\times 2}$, and the distinct state pairs P_p for $p = 1, 2$ are given by $P_1 = [1, 2]$ and $P_2 = [2, 3]$. Then $\mathbf{Q}$ takes the form

$$\mathbf{Q} = \begin{pmatrix} M_{11} - M_{21} - (M_{13} - M_{23}) & M_{12} - M_{22} - (M_{13} - M_{23}) \\ M_{21} - M_{31} - (M_{23} - M_{33}) & M_{22} - M_{32} - (M_{23} - M_{33}) \end{pmatrix} \tag{D52}$$

and its inverse can hence be easily computed:

$$\mathbf{Q}^{-1} = \frac{1}{\det \mathbf{Q}} \begin{pmatrix} M_{22} - M_{32} - (M_{23} - M_{33}) & (M_{13} - M_{23}) - (M_{12} - M_{22}) \\ (M_{23} - M_{33}) - (M_{21} - M_{31}) & M_{11} - M_{21} - (M_{13} - M_{23}) \end{pmatrix} \tag{D53}$$

Using Eq. (D49), we can therefore write for the nontrivial TPP-learned filters

$$\begin{aligned} \vec{f}_1 = \frac{\mathbf{V}^{-1}}{\det \mathbf{Q}} \Big\{ & [M_{22} - M_{32} - (M_{23} - M_{33})]\,\vec{s}^{(1)} + [M_{13} - M_{12} - (M_{33} - M_{32})]\,\vec{s}^{(2)} \\ & + [M_{12} - M_{22} - (M_{13} - M_{23})]\,\vec{s}^{(3)} \Big\}, \end{aligned} \tag{D54a}$$

$$\begin{aligned} \vec{f}_2 = \frac{\mathbf{V}^{-1}}{\det \mathbf{Q}} \Big\{ & [M_{23} - M_{33} - (M_{21} - M_{31})]\,\vec{s}^{(1)} + [M_{11} - M_{13} - (M_{31} - M_{33})]\,\vec{s}^{(2)} \\ & + [M_{13} - M_{23} - (M_{11} - M_{21})]\,\vec{s}^{(3)} \Big\}. \end{aligned} \tag{D54b}$$

Note that the final filter $\vec{f}_3$ must be defined by the constraint expressed by Eq. (16) [or equivalently Eq. (D25)].

6. TPP-learned optimal filters for multistate classification under Gaussian white noise

We now present an example of TPP-learned optimal filters for dispersive qubit readout where the dominant noise source is additive Gaussian white noise. This is ensured via a theoretical simulation of Eq. (6) as discussed in Sec. III A. These simulations yield single-shot measurement records for any number of transmon states. Examples of these records are shown in Fig. 10 for four distinct transmon states $p \in \{e, g, f, h\}$; for ease of visualization, we consider only the I quadrature. We use this simulated dataset as a training set to determine the TPP-learned filters under the white noise assumption, as defined by Eq. (14) with $\mathbf{V} \propto \bar{\mathbf{I}}$. While the individual measurement records are obscured by white noise, the empirically calculated mean traces at the top right in Fig. 10 illustrate the physics at play. The mean traces grow once the measurement tone is turned on past $\mathcal{T}_{\text{on}}$ and settle to a steady state depending on the induced dispersive shift χ_p and the measurement amplitude. The traces begin to fall beyond $\mathcal{T}_{\text{off}}$ and eventually settle to background levels. These means, together with an estimate of the variances, determine the coefficients C_{kp} that define the contribution of the mean trace $\vec{s}^{(p)}$ to the k th filter, and are hence sufficient to calculate optimal filters for the classification of any subset of states.

For the standard binary classification task ($C = 2$) of distinguishing $\{e, g\}$ states, the learned filters are represented in black in the top row in Fig. 10, together with bar plots showing the coefficients C_{kp}. Again for visualization, we show filters $\vec{f}_k \in \mathbb{R}^{N_T}$ only for I-quadrature data; the

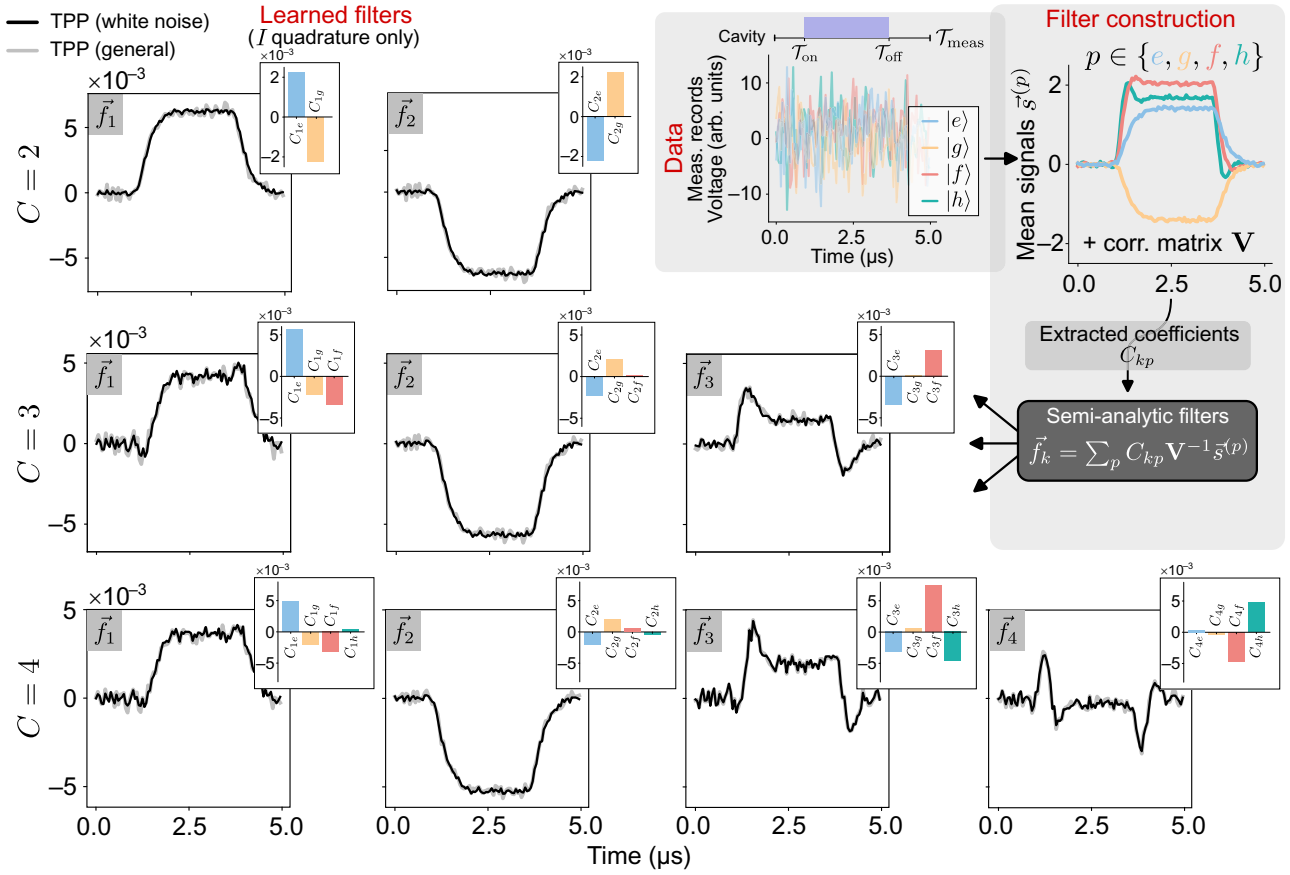


FIG. 10. TPP-learned optimal filters for simulated multistate classification under Gaussian white noise conditions. Top right: Single-shot measurement records obtained under the indicated measurement tone, and empirical mean traces of several heterodyne records of the cavity I quadrature corresponding to multilevel atom states $|p\rangle$, where $p \in \{e,g,f,h\}$. For a transmon $\chi_p/\kappa \in \{-\chi, \chi, -3\chi, -5\chi\}$, $\chi/\kappa = 0.195$, and $\kappa/2\pi = 1.54$ MHz. Rows: TPP-learned optimal filters for classifying states $p \in \{e,g\}$ ($C = 2$), $\{e,g,f\}$ ($C = 3$), and $\{e,g,f,h\}$ ($C = 4$). Black curves represent filters learned under the white noise assumption, calculated analytically with Eq. (14). Bar plots show the coefficients C_{kp} applied to respective mean traces in calculating these filters. Gray curves represent general filters calculated by our numerically solving Eq. (2). Analytically computed white noise filters and general filters can both be extended to arbitrary C.

complete vector $\vec{\boldsymbol{f}}_k$ includes filters for all N_O observables. For the binary case, the $k = 1$ TPP-learned filter *always* satisfies $C_{1e} = -C_{1g}$. Hence, it is simply proportional to the difference of mean traces for the two states, $\vec{f}_1 \propto \vec{s}^{(e)} - \vec{s}^{(g)}$, making it exactly equivalent to the standard matched filter for binary classification. We note that the second filter ($k = 2$) is simply the negative of the first, as demanded by Eq. (16).

Crucially, the TPP approach now provides the generalization of such matched filters to the classification of an arbitrary number of states. For three-state ($C = 3$) classification of $\{e,g,f\}$ states, the three TPP-learned filters are plotted in the middle row, while the last row shows the four filters for the classification of $C = 4$ states $\{e,g,f,h\}$. Filters for the classification of an arbitrary number of states C can be constructed similarly. The bar plots of C_{kp} show how these filters typically have nonzero contributions from the mean traces for *all* states. This emphasizes that the TPP-learned filters are not simply a collection of binary matched filters but are a more nontrivial construction. Most importantly, our analytic approach enables this construction by inverting a matrix in $\mathbb{R}^{(C-1)\times(C-1)}$ to determine C_{kp}. This has a substantially lower complexity relative to the pseudoinverse calculation demanded by Eq. (2), which requires inversion of a much larger matrix $\mathbf{C} \in \mathbb{R}^{N_O N_T \times N_O N_T}$.

Of course, the latter approach of obtaining $\mathbf{W}^{\text{opt}}$ and hence TPP filters using Eq. (2) can also be applied for learning using the same training data. Here it yields the underlying filters represented in gray. The resulting filters appear to simply be noisier versions of the analytically calculated filters. The reason for this straightforward: the fact that the noise in the measurement data is additive Gaussian white noise is a key piece of information used in calculating the white noise TPP filters, but is not *a priori* known to the general TPP. The latter makes no

assumptions regarding the underlying noise statistics of the dataset. Instead, the training procedure itself enables the TPP to learn the statistics of the noise and adjust $\mathbf{W}^{\text{opt}}$ accordingly. The fact that the temporal profile of general TPP filters gradually approaches that of the white noise filters shows this learning in practice. This ability to extract noise statistics from data is a key feature that makes TPP learning useful under more general noise conditions, as demonstrated in Secs. IV and V.

7. Comparison of the TPP against $C-1$ instances of FGDA

In Sec. III A, the performance of TPP-learned optimal filters was compared against standard FGDA implementations where a single matched filter is used. However, as the TPP uses $C-1$ independent filters, it is natural to ask for $C>2$ state classification tasks whether use of multiple instances of FGDA with distinct filters could provide an improvement in performance. In other words, does the improvement in TPP performance observed in Fig. 2 arise simply because the TPP is using more filters or is due to the learned optimal filters being able to extract more useful information from the noisy temporal measurement data?

To investigate this, we consider the $C=3$ state classification task from Sec. III A, but now compare the TPP against $C-1$ instances of the FGDA. A standard approach to do so is to consider one-versus-all classification. Here, for a single instance, an FGDA is trained to process temporal data and to output a state label as being p, or *not* p (or $!p$ for short), instead of predicting a precise state label in the $!p$ case. The "filter" portion of this FGDA can be labeled a one-versus-all matched filter, and can be constructed, for example, as

$$\vec{h}_{I,p}=\frac{1}{N_{\text{train}}}\sum_{n=1}^{N_{\text{train}}}\left(\vec{I}_{(n)}^{(p)}-\frac{1}{C-1}\sum_{p'\neq p}\vec{I}_{(n)}^{(p')}\right). \tag{D55}$$

Next, a second instance of the FGDA processes the same temporal data but using a one-versus-all matched filter constructed for a *different* state label q, and hence now predicts the state label as being q or $!q$. FGDA instances are used to process temporal data until $C-1$ instances have been used and hence one of $2^{(C-1)}$ possible outcomes has been obtained. A concrete example of the possible outcomes for $C=3$ state classification is shown in Fig. 11(a) for one-versus-all filters constructed for $p=g$ and $q=e$. Depending on the possible joint outcome, a state label can finally be assigned: for example, the result g and $!e$ is consistent with the state label g, $!g$ and e implies e, and $!g$ and $!e$ implies f. Note that the final outcome g and e is ambiguous; here we use a random choice to assign a state label.

Note that use of $C-1$ instances of the FGDA introduces more ambiguities than use of just a single filter: different choices of p and q can be made, as indicated by the other tables in Fig. 11(a), where we choose either $p=e$, $q=f$ or $p=g$, $q=f$. There is even greater ambiguity about the choice of the $C-1$ one-versus-all matched filters, Eq. (D55), where the prefactors of each mean trace can be chosen arbitrarily. Even before exploring the performance of $C-1$ FGDA instances, we note that the TPP already provides a unique set of filters, determined by coefficients C_{kp} as given by Eq. (D49).

We now compare the performance of the TPP against the three distinct $C-1$ FGDA instance implementations shown in Fig. 11(a), first for the readout conditions from Fig. 2; the results are shown in Fig. 11(b). Also shown is the use of a single g-f matched filter, which was found to match the performance of the TPP in this case. We clearly see that the performance of the three distinct $C-1$ FGDA instances matches the TPP performance much more closely than the single matched filters used in Fig. 2. However, if the readout conditions are modified, for example, if the cavity readout drive is now resonant with the cavity frequency when the qubit is in the ground state $|g\rangle$, the performance can vary significantly, as shown in Fig. 11(c). All $C-1$ FGDA instances have a higher classification infidelity than the TPP, with certain instances faring much worse than others.

It is therefore clear that the improvement in classification fidelity provided by the TPP is *not* due only to its use of more than a single filter: $C-1$ FGDA instances using the same number of independent filters as the TPP do not always match its performance. This emphasizes the need to optimize the individual filters used; the TPP provides an autonomous, model-free approach to achieve precisely this objective for the classification of an arbitrary number of states.

8. Semianalytic TPP-learned optimal filters beyond Gaussian white noise conditions

As shown in the main text, a key feature of the TPP is that it applies to postprocessing of temporal data experiencing more general noise conditions than simply uniform, observable-independent Gaussian white noise. In the main text, we compared numerically calculated general TPP filters with semianalytic filters computed under the white noise approximation. In this subsection we also show the semianalytic but general TPP filters, as defined by Eq. (14) for a general correlation matrix $\mathbf{V}$.

We start with a simple case where the required $\mathbf{V}^{-1}$ can be computed analytically. Consider the case of heterodyne measurement $N_{\text{O}}=2$ but where the two measured observables (quadrature time series $\vec{I}$ and $\vec{Q}$) have stationary but *distinct* variances σ_I^2 and σ_Q^2 respectively; for concreteness,

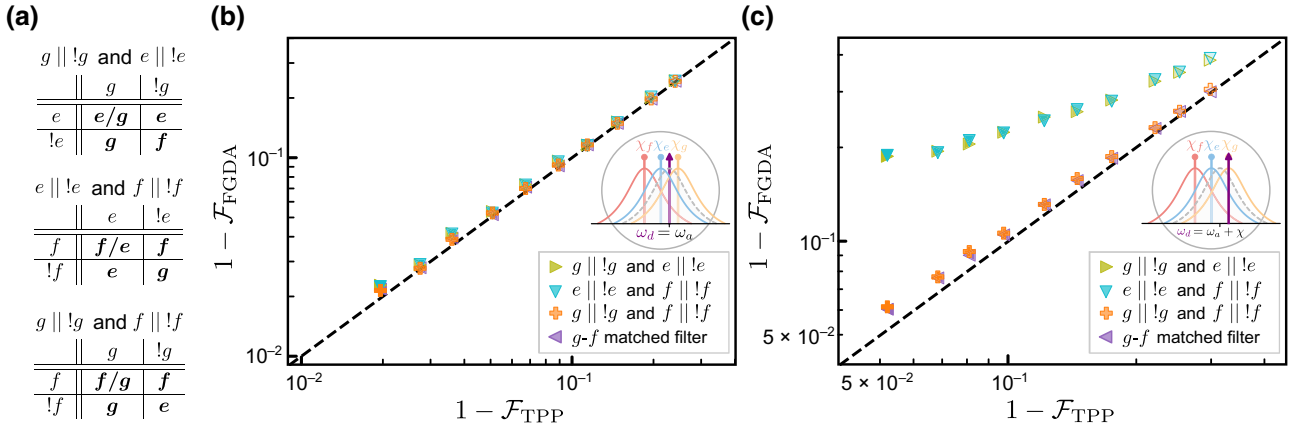


FIG. 11. Multistate ($C = 3$) classification performance of the TPP versus $C - 1$ instances of FGDA under Gaussian white noise conditions. (a) Three distinct schemes (each corresponding to a different table) to implement $C = 3$ state classification using $C - 1$ instances of FGDA. Each instance predicts outcomes given by headers of rows and columns, respectively, while bold labels indicate final predicted labels based on joint outcomes; see discussion in Appendix D 7 for details. Performance comparison for (b) the same readout conditions as for Fig. 2 and (c) for readout conditions where the measurement drive is resonant with the dispersively shifted cavity when the transmon qubit is in state $|g\rangle$. The TPP still outperforms $C - 1$ FGDA instances, with the latter's performance also varying depending on the readout conditions.

we assume $\sigma_Q^2 > \sigma_I^2$. In this case, $\mathbf{V}$ takes the simple form

$$\mathbf{V} = \begin{pmatrix} \sigma_I^2 \tilde{\mathbf{I}} & \mathbf{0} \\ \mathbf{0} & \sigma_Q^2 \tilde{\mathbf{I}} \end{pmatrix}, \tag{D56}$$

where $\tilde{\mathbf{I}}$ is the identity matrix in $\mathbb{R}^{N_T \times N_T}$. Of course, this form of $\mathbf{V}$ can be straightforwardly inverted:

$$\mathbf{V}^{-1} = \begin{pmatrix} \frac{1}{\sigma_I^2} \tilde{\mathbf{I}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma_Q^2} \tilde{\mathbf{I}} \end{pmatrix}. \tag{D57}$$

For convenience, we define filters and mean traces for each quadrature as $\vec{f}_k = \begin{pmatrix} \vec{f}_k^{\,I} \\ \vec{f}_k^{\,Q} \end{pmatrix}$ and $\vec{s}_k^{(p)} = \begin{pmatrix} \vec{s}_k^{I(p)} \\ \vec{s}_k^{Q(p)} \end{pmatrix}$, respectively. To calculate the semianalytic general filters, we then simply use Eq. (14) to immediately find

$$\begin{aligned} \vec{f}_k^{\,I} &= \sum_k C_{kp}(\mathbf{V}) \frac{1}{\sigma_I^2} \vec{s}_k^{I(p)}, \\ \vec{f}_k^{\,Q} &= \sum_k C_{kp}(\mathbf{V}) \frac{1}{\sigma_Q^2} \vec{s}_k^{Q(p)}. \end{aligned} \tag{D58}$$

We see that there is now a relative weighting of the filters in accordance with their variance: noisier observables are suppressed relative to less noisy observables. Additionally, the coefficients C_{kp} also depend on $\mathbf{V}^{-1}$. In Fig. 12(a) we plot the resulting filters for the readout conditions considered in Fig. 11(b) (this ensures the I and Q quadratures both have nonzero mean signal values) for both the semianalytic general TPP filters given by Eq. (D58) and the

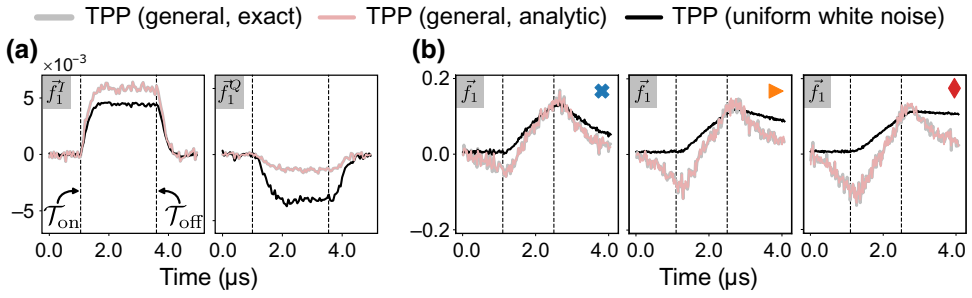


FIG. 12. Comparison of semianalytic and exact TPP filters under general noise correlation conditions. Filters are shown for (a) simulated data where different quadrature time series $\vec{I}$, $\vec{Q}$ have different variances σ_I^2, σ_Q^2, respectively, and (b) simulated data from a phase-preserving quantum-limited amplifier, as in Sec. V A. Excellent agreement is observed between the exact general TPP filter and the semianalytic general TPP filter, while both are markedly different from TPP filters under the assumption of uniform (namely, observable-independent) Gaussian white noise, as represented by the black curves.

exact general TPP filters; the latter are shown with thicker lines deliberately to highlight differences between the two (to be expanded upon in due course). Finally, also shown are the filters with the assumption of uniform white noise across the measured quadratures, which are clearly distinct from the general filters and do not penalize the noisier Q quadrature.

Secondly, in Fig. 12(b), we consider the case of correlated quantum noise added by a finite-bandwidth phase-preserving quantum amplifier from Sec. V A, now also showing calculated semianalytic general filters. We see that for all cases the semianalytic general TPP filters show only very small differences when compared with the exact general TPP filter. As both schemes use empirically calculated mean traces to construct the Gram matrix $\mathbf{G}$ and empirically estimate the correlation matrix $\mathbf{V}$, the residual differences can be attributed to the fact that the semianalytic TPP filter assumes the noise terms have zero mean, while the exact general filter does not make such an assumption.

We also emphasize that computing the exact general TPP filter requires the inversion of the matrix $\mathbf{C} \in \mathbb{R}^{(N_O N_T+1)\times(N_O N_T+1)}$, while the semianalytic general TPP requires the inversion of $\mathbf{V} \in \mathbb{R}^{(N_O N_T)\times(N_O N_T)}$. In the general case, the numerical advantage in inverting the slightly smaller matrix $\mathbf{V}$ is not as significant as it is in the special case where $\mathbf{V}$ is proportional to the identity matrix. However, in cases where an analytic form for $\mathbf{V}$ and more importantly its inverse is similarly known, the semianalytic general TPP filter can be more numerically efficient than calculating these filters exactly.

APPENDIX E: SUPPLEMENTARY CLASSIFICATION RESULTS

1. Classification performance versus increasing signal amplitude for real qubit readout

In theory, namely, ignoring measurement chain nonidealities and qubit transitions discussed in Sec. III, increase in measurement tone amplitude leads to an increase in qubit classification fidelity. However, for real qubits, additional effects can create complex readout conditions such that increasing the measurement tone amplitude may not uniformly increase readout fidelity. To analyze whether increased readout power facilitates increased readout fidelity for the real qubit readout data collected in this work, we analyze the data in Fig. 3 in a slightly different form. We first introduce the quantity

$$\mathcal{N}_j(s) = \frac{1-\mathcal{F}_j(s)}{1-\mathcal{F}_j(s_0)}, \tag{E1}$$

where $j \in \{\mathrm{FGDA}, \mathrm{TPP}\}$ depending on the classification scheme used, while s denotes signal amplitude and s_0 is the smallest signal amplitude for a given dataset. As a result, $\mathcal{N}_j$ is simply the infidelity as a function of measurement tone amplitude, normalized by the infidelity at the smallest amplitude; we therefore require $\mathcal{N}_j < 1$ for a reduction in readout *in*fidelity (and hence an increase in readout *fidelity*) for increasing readout power.

In Fig. 13, we plot $\mathcal{N}_{\mathrm{FGDA}}$ against $\mathcal{N}_{\mathrm{TPP}}$ for the two dispersive qubit-cavity systems that were analyzed as a function of measurement tone amplitude in the main text. The data point at $\mathcal{N}_{\mathrm{FGDA}} = \mathcal{N}_{\mathrm{TPP}} = 1$ for each dataset corresponds to the lowest amplitude, by construction of

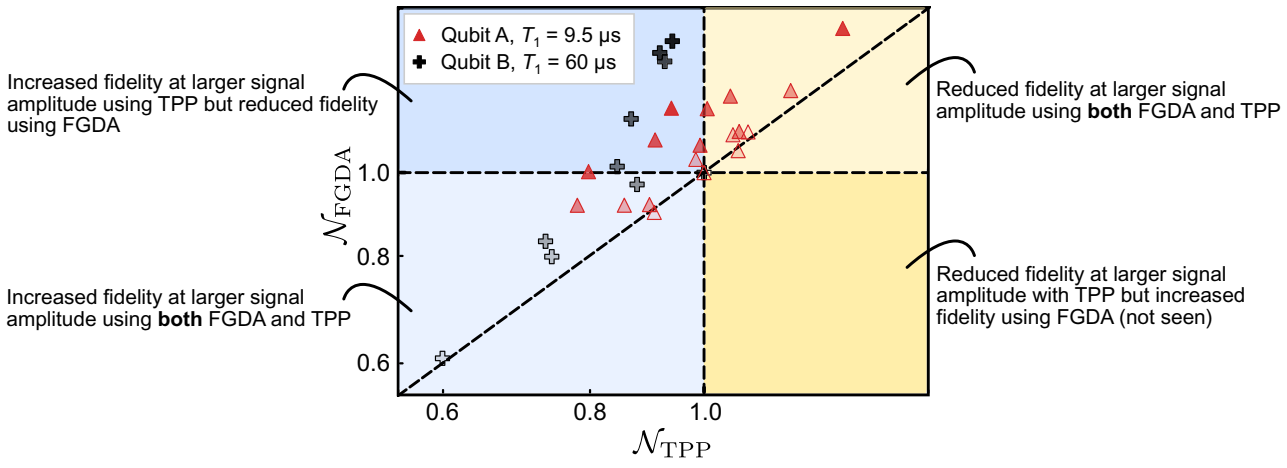


FIG. 13. Classification performance of the TPP versus FGDA as a function of increasing measurement tone (signal) amplitude for readout of real qubits. Same data as for Fig. 3, but now with our plotting $\mathcal{N}$, the classification infidelity normalized by infidelity at the lowest signal amplitude for each dispersive qubit-cavity system shown; see Eq. (E1). More-opaque markers indicate higher measurement tone amplitudes.

Eq. (E1). We next note that a few data points fall into the category where both $\mathcal{N}_{\mathrm{TPP}} > 1$ and $\mathcal{N}_{\mathrm{FGDA}} > 1$, indicating increased infidelity with increasing signal amplitude with either classification scheme. Here nonidealities such as enhanced qubit transitions degrade the measured data, which neither classification scheme is able to overcome. We note that here we still have $\mathcal{N}_{\mathrm{FGDA}} > \mathcal{N}_{\mathrm{TPP}}$, so the FGDA performance is worse than the TPP performance.

Of all the other data points corresponding to higher readout amplitudes, most lie in the blue shaded regions of the plot, where $\mathcal{N}_{\mathrm{TPP}} < 1$. These are qubit readout conditions for which increasing the measurement tone amplitude leads to improved classification performance when the TPP is used. Of these points, just over half also have $\mathcal{N}_{\mathrm{FGDA}} < 1$, implying that use of either the FGDA or the TPP provides an improvement. Again, $\mathcal{N}_{\mathrm{FGDA}} > \mathcal{N}_{\mathrm{TPP}}$, so the improvement is larger with the TPP.

Crucially, the other half of the data points are such that $\mathcal{N}_{\mathrm{TPP}} < 1$ while $\mathcal{N}_{\mathrm{FGDA}} > 1$. For these regimes, the use of the TPP is necessary to extract an advantage in readout fidelity with increasing signal amplitude. Equally as importantly, *none* of the data points fall in the category where $\mathcal{N}_{\mathrm{TPP}} > 1$ while $\mathcal{N}_{\mathrm{FGDA}} < 1$; this indicates that there is, in general, no disadvantage in deploying the TPP instead of the FGDA at higher signal amplitudes, as the TPP will not be outperformed by the FGDA. Together, these results supplement findings in the main text that the TPP can provide a robust classification scheme to extract maximum performance in complex readout regimes at high powers.

2. Three-state classification results for real qubit readout

In this section we include some results supplementary to Fig. 3, now comparing classification performance for multistate ($C = 3$) classification for real qubit readout of $p \in \{e, g, f\}$. The results are shown in Fig. 14 for the readout of qubit B.

The standard FGDA is deployed here with use of the g-f matched filter, introduced in Sec. III; as discussed in Appendix D 7, this provides the best performance among other single matched filters, while $C - 1$ matched filters do not provide a marked improvement in this readout configuration. We note again that the TPP outperforms the FGDA for almost all data points, and the performance difference increases at higher measurement tone amplitudes. The underperformance at the lowest measurement tone amplitude can again be attributed to the fact that under these simpler readout conditions, the temporal profile of the optimal filter is close to that of the white noise filter (see Fig. 4); the general TPP does not make any assumptions about the noise statistics *a priori*, and must learn these from a finite training dataset, whose size limits constrains the fidelity of the learned filter. At higher signal amplitudes, the TPP outperforms the FGDA in spite of this training cost. Overall, we see that the TPP provides a better classification scheme for multistate readout of real qubits, supplementing the improvement in performance demonstrated for binary classification of real qubits in the main text.

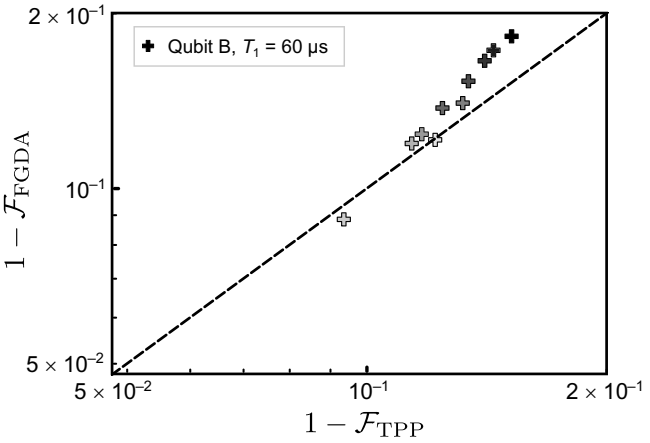


FIG. 14. Multistate ($C = 3$) classification performance of the TPP versus FGDA for readout of real qubits. Classification infidelities obtained with both schemes are plotted against each other for one of the three dispersive qubit-cavity systems analyzed in the main text. The dashed line marks $1 - \mathcal{F}_{\mathrm{FGDA}} = 1 - \mathcal{F}_{\mathrm{TPP}}$. More-opaque markers indicate higher measurement tone amplitudes.

3. TPP learning of correlated *classical* noise

In this section, we use a further example to demonstrate the ability of TPP-based learning to extract correlations from measured data to supplement simulations in Sec. V. As in Sec. V A, we again consider simulated datasets of measured heterodyne records from a measurement chain of a qubit-cavity-amplifier setup, as in Appendix D 6. Now, however, we consider the excess classical noise added by the measurement process to also possess a component with a colored spectrum (suppressing quadrature labels for clarity):

$$\xi^{\mathrm{cl}}(t_i) = \sigma^{\mathrm{W}}\xi^{\mathrm{W}}(t_i) + \sigma^{\mathrm{P}}\xi^{\mathrm{P}}(t_i), \tag{E2}$$

where $\xi^{\mathrm{W}}(t_i)$ describes white noise as before, while $\xi^{\mathrm{P}}(t_i)$ describes $1/f$ (or pink) noise. The power spectral density of the noise processes is given by the Fourier transform of their steady-state autocorrelation function (by the Wiener-Khinchin theorem), $S_{\mathrm{N}}[f] = \int d\tau\, e^{-i2\pi f\tau}\mathbb{E}[\xi^{\mathrm{N}}(0)\xi^{\mathrm{N}}(\tau)]$ for $N \in \{W, P\}$. The noise processes are normalized so that the total noise power $\int df\ |S_{\mathrm{N}}[f]|$ is the same for any of the noise processes considered; hence, the relative magnitude $(\sigma^{\mathrm{P}}/\sigma^{\mathrm{W}})^2$ determines the relative strength of the noise processes with different correlation statistics.

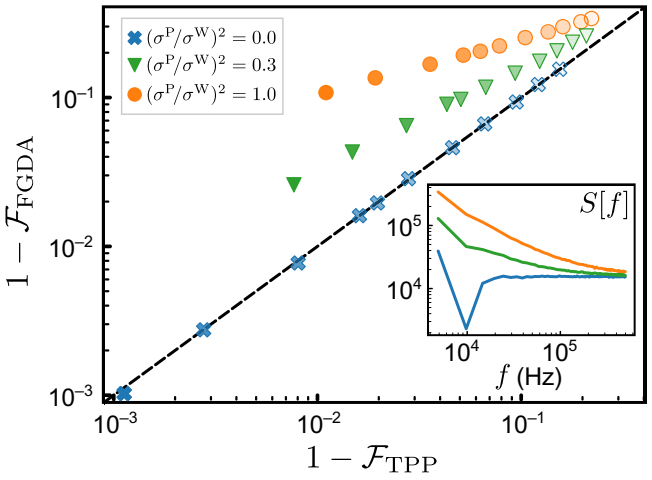


FIG. 15. Comparative classification performance of FGDA versus the TPP in the presence of classical correlated noise. We consider a $C = 2$ (binary) dispersive qubit readout task using simulated data and for different colored noise conditions. Darker markers indicate higher measurement tone amplitudes. The dashed line indicates $1 - \mathcal{F}_{\mathrm{FGDA}} = 1 - \mathcal{F}_{\mathrm{TPP}}$. The inset shows the corresponding noise spectral density $S[f]$, which remains unchanged with coherent input power.

We restrict ourselves again to binary classification of states $|e\rangle$ and $|g\rangle$. In Fig. 15, we plot the calculated infidelities obtained with the MF and TPP approaches against each other on a logarithmic scale for different noise conditions parameterized by $(\sigma^{\mathrm{P}}/\sigma^{\mathrm{W}})^2$ and as a function of the coherent input tone power: darker markers correspond to readout with stronger input tones.

We immediately see that if the excess classical noise is purely white noise, the FGDA and the TPP exhibit very similar performance: both lie along the dashed line of equal infidelities. However, the situation is very different if the added noise is colored noise, namely, $(\sigma^{\mathrm{P}}/\sigma^{\mathrm{W}})^2 \neq 0$, and hence has a nonzero correlation timescale. We immediately note that even when the colored noise power is only a fraction of the white noise power, the TPP-learned filters provide a non-negligible improvement over the standard FGDA scheme using matched filters.

APPENDIX F: TIME-SHUFFLED DATA

As discussed in Appendix C, the trained weights $\mathbf{W}$ take the form of Eq. (C8),

$$\mathcal{W}^{\mathrm{opt}} = \mathbf{Y}\mathcal{X}^T(\mathcal{X}\mathcal{X}^T - \lambda\mathbf{I})^{-1}. \tag{F1}$$

We now consider the operation of a matrix $\mathbf{J}$ on $\mathbf{X}$ that serves to reorder the time indices of measurement records; this amounts to an exchange of specific rows of $\mathbf{X}$ and is therefore referred to as an exchange matrix, a special case of the more general permutation matrix in standard linear algebra. As $\mathcal{X} \in \mathbb{R}^{(N_{\mathrm{O}}N_{\mathrm{T}}+1)\times CN_{\mathrm{train}}}$ and the exchange matrix is intended to switch *rows* of the data, we must have $\mathbf{J} \in \mathbb{R}^{(N_{\mathrm{O}}N_{\mathrm{T}}+1)\times(N_{\mathrm{O}}N_{\mathrm{T}}+1)}$. Furthermore, the exchange matrix satisfies the properties $\mathbf{J}^{-1} = \mathbf{J} = \mathbf{J}^T$, so $\mathbf{JJ} = \mathbf{I}$.

We therefore define a new data matrix $\mathcal{X}_J$ with exchanged rows under the action of the exchange matrix:

$$\mathcal{X}_J = \mathbf{J}\mathcal{X} \implies \mathcal{X} = \mathbf{J}\mathcal{X}_J, \tag{F2}$$

where we have used the property that $\mathbf{J}^{-1} = \mathbf{J}$. Note that the target matrix $\mathbf{Y}$ is unchanged, since the particular class a measurement record belongs to should not be related to time ordering of the measurement records.

The trained weights can equivalently be written as

$$\mathcal{W}^{\mathrm{opt}} = \mathbf{Y}(\mathbf{J}\mathcal{X}_J)^T(\mathbf{J}\mathcal{X}_J\mathcal{X}_J^T\mathbf{J}^T - \lambda\mathbf{I})^{-1}, \tag{F3}$$

which, after some simplification and the use of $\mathbf{J}^T = \mathbf{J}$, reduces to

$$\begin{aligned}\mathcal{W}^{\mathrm{opt}} &= \mathbf{Y}\mathcal{X}_J^T\mathbf{JJ}(\mathcal{X}_J\mathcal{X}_J^T)^{-1}\mathbf{J} \\ &= \left[\mathbf{Y}\mathcal{X}_J^T(\mathcal{X}_J\mathcal{X}_J^T - \lambda\mathbf{I})^{-1}\right]\mathbf{J}.\end{aligned} \tag{F4}$$

The term in square brackets is simply the new trained weights when the exchanged data matrix $\mathcal{X}_J$ is used; we label this as $(\mathcal{W}^{\mathrm{opt}})_J$. We therefore find

$$(\mathcal{W}^{\mathrm{opt}})_J = \mathcal{W}^{\mathrm{opt}}\mathbf{J}, \tag{F5}$$

which simply indicates that the new trained weights are simply exchanged versions of the previous trained weights.

[1] A. Roy and M. Devoret, Introduction to parametric amplification of quantum signals with Josephson circuits, C. R. Phys. **17**, 740 (2016).

[2] J. Aumentado, Superconducting parametric amplifiers: The state of the art in Josephson parametric amplifiers, IEEE Microw. Mag. **21**, 45 (2020).

[3] A. P. M. Place, *et al.*, New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds, Nat. Commun. **12**, 1779 (2021).

[4] G. Angelatos, S. A. Khan, and H. E. Türeci, Reservoir computing approach to quantum state measurement, Phys. Rev. X **11**, 041062 (2021).

[5] S. A. Khan, F. Hu, G. Angelatos, and H. E. Türeci, Physical reservoir computing using finitely-sampled quantum systems, ArXiv:2110.13849.

[6] J. Nokkala, R. Martínez-Peña, G. L. Giorgi, V. Parigi, M. C. Soriano, and R. Zambrini, Gaussian states of continuous-variable quantum systems provide universal and versatile reservoir computing, Commun. Phys., **4**, 1 (2021).

[7] R. Martínez-Peña, G. L. Giorgi, J. Nokkala, M. C. Soriano, and R. Zambrini, Dynamical phase transitions in quantum reservoir computing, Phys. Rev. Lett. **127**, 100502 (2021).

[8] P. Mujal, R. Martínez-Peña, J. Nokkala, J. García-Beni, G. L. Giorgi, M. C. Soriano, and R. Zambrini, Opportunities in quantum reservoir computing and extreme learning machines, Adv. Quantum Technol., **4**, 2100027 (2021).
[9] D. Sank, *et al.*, Measurement-induced state transitions in a superconducting qubit: Beyond the rotating wave approximation, Phys. Rev. Lett. **117**, 190503 (2016).
[10] M. Malekakhlagh, A. Petrescu, and H. E. Türeci, Lifetime renormalization of weakly anharmonic superconducting qubits. I. Role of number nonconserving terms, Phys. Rev. B **101**, 134509 (2020).
[11] A. Petrescu, M. Malekakhlagh, and H. E. Türeci, Lifetime renormalization of driven weakly anharmonic superconducting qubits. II. The readout problem, Phys. Rev. B **101**, 134510 (2020).
[12] R. Hanai, A. McDonald, and A. Clerk, Intrinsic mechanisms for drive-dependent Purcell decay in superconducting quantum circuits, Phys. Rev. Res. **3**, 043228 (2021).
[13] M. Khezri, A. Opremcak, Z. Chen, K. C. Miao, M. McEwen, A. Bengtsson, T. White, O. Naaman, D. Sank, A. N. Korotkov, Y. Chen, and V. Smelyanskiy, Measurement-induced state transitions in a superconducting qubit: Within the rotating-wave approximation, Phys. Rev. Appl. **20**, 054008 (2023).
[14] R. Shillito, A. Petrescu, J. Cohen, J. Beall, M. Hauru, M. Ganahl, A. G. Lewis, G. Vidal, and A. Blais, Dynamics of transmon ionization, Phys. Rev. Appl. **18**, 034031 (2022).
[15] T. Thorbeck, Z. Xiao, A. Kamal, and L. C. Govia, Readout-induced suppression and enhancement of superconducting qubit lifetimes, Phys. Rev. Lett. **132**, 090602 (2024).
[16] D. Gusenkova, M. Spiecker, R. Gebauer, M. Willsch, D. Willsch, F. Valenti, N. Karcher, L. Grünhaupt, I. Takmakov, P. Winkel, D. Rieger, A. V. Ustinov, N. Roch, W. Wernsdorfer, K. Michielsen, O. Sander, and I. M. Pop, Quantum nondemolition dispersive readout of a superconducting artificial atom using large photon numbers, Phys. Rev. Appl. **15**, 064030 (2021).
[17] M. F. Dumas, B. Groleau-Paré, A. McDonald, M. H. Muñoz-Arias, C. Lledó, B. D'Anjou, and A. Blais, Unified picture of measurement-induced ionization in the transmon, ArXiv:2402.06615v1.
[18] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, Rapid high-fidelity single-shot dispersive readout of superconducting qubits, Phys. Rev. Appl. **7**, 054020 (2017).
[19] M. Tsang, Volterra filters for quantum estimation and detection, Phys. Rev. A **92**, 062119 (2015).
[20] B. Lienhard, A. Vepsäläinen, L. C. Govia, C. R. Hoffer, J. Y. Qiu, D. Ristè, M. Ware, D. Kim, R. Winik, A. Melville, B. Niedzielski, J. Yoder, G. J. Ribeill, T. A. Ohki, H. K. Krovi, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Deep-neural-network discrimination of multiplexed superconducting-qubit states, Phys. Rev. Appl. **17**, 014024 (2022).
[21] J. Gambetta, W. A. Braff, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, Protocols for optimal readout of qubits using a continuous quantum nondemolition measurement, Phys. Rev. A **76**, 012325 (2007).
[22] Source code available at https://zenodo.org/doi/10.5281/zenodo.10020462.
[23] I. Takmakov, P. Winkel, F. Foroughi, L. Planat, D. Gusenkova, M. Spiecker, D. Rieger, L. Grünhaupt, A. Ustinov, W. Wernsdorfer, I. Pop, and N. Roch, Minimizing the discrimination time for quantum states of an artificial atom, Phys. Rev. Appl. **15**, 064029 (2021).
[24] Y. Sunada, S. Kono, J. Ilves, S. Tamate, T. Sugiyama, Y. Tabuchi, and Y. Nakamura, Fast readout and reset of a superconducting qubit coupled to a resonator with an intrinsic Purcell filter, Phys. Rev. Appl. **17**, 044016 (2022).
[25] A. Bengtsson, A. Opremcak, M. Khezri, D. Sank, A. Bourassa, K. J. Satzinger, S. Hong, C. Erickson, B. J. Lester, K. C. Miao, A. N. Korotkov, J. Kelly, Z. Chen, and P. V. Klimov, Model-based optimization of superconducting qubit readout, Phys. Rev. Lett. **132**, 100603 (2024).
[26] D. Sank, A. Opremcak, A. Bengtsson, M. Khezri, Z. Chen, O. Naaman, and A. Korotkov, System characterization of dispersive readout in superconducting qubits, ArXiv:2402.00413.
[27] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, Recent advances in physical reservoir computing: A review, Neural Netw. **115**, 100 (2019).
[28] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa, Next generation reservoir computing, Nat. Commun., **12**, 5564 (2021).
[29] P. Luchi, P. E. Trevisanutto, A. Roggero, J. L. DuBois, Y. J. Rosen, F. Turro, V. Amitrano, and F. Pederiva, Enhancing qubit readout with autoencoders, Phys. Rev. Appl. **20**, 014045 (2023).
[30] S. Maurya, C. N. Mude, W. D. Oliver, B. Lienhard, and S. Tannu, in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, ISCA '23 (Association for Computing Machinery, New York, NY, USA, 2023), p. 1.
[31] D. Canaday, A. Griffith, and D. J. Gauthier, Rapid time series prediction with a hardware-based reservoir computer, Chaos **28**, 123119 (2018).
[32] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data, Chaos **27**, 121102 (2017).
[33] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Phys. Rev. Lett. **120**, 024102 (2018).
[34] A. Griffith, A. Pomerance, and D. J. Gauthier, Forecasting chaotic systems with very low connectivity reservoir computers, Chaos **29**, 123108 (2019).
[35] D. Canaday, A. Pomerance, and D. J. Gauthier, Model-free control of dynamical systems with deep reservoir computing, J. Phys.: Complex. **2**, 035025 (2021).
[36] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, Phys. Rev. A **93**, 060302 (2016).
[37] J. Kelly, P. O. Malley, M. Neeley, H. Neven, and J. M. M. Google, Physical qubit calibration on a directed acyclic graph, ArXiv:1803.03226.

[38] X. Dai, D. M. Tennant, R. Trappen, A. J. Martinez, D. Melanson, M. A. Yurtalan, Y. Tang, S. Novikov, J. A. Grover, S. M. Disseler, J. I. Basham, R. Das, D. K. Kim, A. J. Melville, B. M. Niedzielski, S. J. Weber, J. L. Yoder, D. A. Lidar, and A. Lupascu, Calibration of flux crosstalk in large-scale flux-tunable superconducting quantum circuits, PRX Quantum **2**, 040313 (2021).

[39] F. Hu, G. Angelatos, S. A. Khan, M. Vives, E. Türeci, L. Bello, G. E. Rowlands, G. J. Ribeill, and H. E. Türeci, Tackling sampling noise in physical systems for machine learning applications: Fundamental limits and eigentasks, Phys. Rev. X **13**, 041020 (2023).

[40] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access **8**, 42200 (2020).

[41] G. Zhu, D. G. Ferguson, V. E. Manucharyan, and J. Koch, Circuit QED with fluxonium qubits: Theory of the dispersive regime, Phys. Rev. B **87**, 024510 (2013).

[42] A. Blais, A. L. Grimsmo, S. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Rev. Mod. Phys. **93**, 025005 (2021).

[43] F. Mallet, F. R. Ong, A. Palacios-Laloy, F. Nguyen, P. Bertet, D. Vion, and D. Esteve, Single-shot qubit readout in circuit quantum electrodynamics, Nat. Phys. **5**, 791 (2009).

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, 2016), 2nd ed.

[45] G. Turin, An introduction to matched filters, IRE Trans. Inf. Theory **6**, 311 (1960).

[46] M. Silveri, E. Zalys-Geller, M. Hatridge, Z. Leghtas, M. H. Devoret, and S. M. Girvin, Theory of remote entanglement via quantum-limited phase-preserving amplification, Phys. Rev. A **93**, 062310 (2016).

[47] P. Kurpiers, P. Magnard, T. Walter, B. Royer, M. Pechal, J. Heinsoo, Y. Salathé, A. Akin, S. Storz, J. C. Besse, S. Gasparinetti, A. Blais, and A. Wallraff, Deterministic quantum state transfer and remote entanglement using microwave photons, Nature **558**, 264 (2018).

[48] B. A. Kochetov and A. Fedorov, Higher-order nonlinear effects in a Josephson parametric amplifier, Phys. Rev. B **92**, 224304 (2015).

[49] S. Boutin, D. M. Toyli, A. V. Venkatramani, A. W. Eddins, I. Siddiqi, and A. Blais, Effect of higher-order nonlinearities on amplification and squeezing in Josephson parametric amplifiers, Phys. Rev. Appl. **8**, 054030 (2017).

[50] D. J. Parker, M. Savytskyi, W. Vine, A. Laucht, T. Duty, A. Morello, A. L. Grimsmo, and J. J. Pla, Degenerate parametric amplification via three-wave mixing using kinetic inductance, Phys. Rev. Appl. **17**, 034064 (2022).

[51] A. Remm, S. Krinner, N. Lacroix, C. Hellings, F. Swiadek, G. J. Norris, C. Eichler, and A. Wallraff, Intermodulation distortion in a Josephson traveling-wave parametric amplifier, Phys. Rev. Appl. **20**, 034027 (2023).

[52] R. Kaufman, T. White, M. I. Dykman, A. Iorio, G. Sterling, S. Hong, A. Opremcak, A. Bengtsson, L. Faoro, J. C. Bardin, T. Burger, R. Gasca, and O. Naaman, Josephson parametric amplifier with Chebyshev gain profile and high saturation, Phys. Rev. Appl. **20**, 054058 (2023).

[53] R. Kaufman, C. Liu, K. Cicak, B. Mesits, M. Xia, C. Zhou, M. Nowicki, D. Pekker, J. Aumentado, and M. Hatridge (to be published).

[54] L. Chen, H. X. Li, Y. Lu, C. W. Warren, C. J. Križan, S. Kosen, M. Rommel, S. Ahmed, A. Osman, J. Biznárová, A. F. Roudsari, B. Lienhard, M. Caputo, K. Grigoras, L. Grönberg, J. Govenius, A. F. Kockum, P. Delsing, J. Bylander, and G. Tancredi, Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier, npj Quantum Inf. **9**, 1 (2023).

[55] D. I. Schuster, A. Wallraff, A. Blais, L. Frunzio, R. S. Huang, J. Majer, S. M. Girvin, and R. J. Schoelkopf, as stark shift and dephasing of a superconducting qubit strongly coupled to a cavity field, Phys. Rev. Lett. **94**, 123602 (2005).

[56] J. Cohen, A. Petrescu, R. Shillito, and A. Blais, Reminiscence of classical chaos in driven transmons, PRX Quantum **4**, 020312 (2023).

[57] M. Khezri, E. Mlinar, J. Dressel, and A. N. Korotkov, Measuring a transmon qubit in circuit QED: Dressed squeezed states, Phys. Rev. A **94**, 012347 (2016).

[58] D. T. McClure, H. Paik, L. S. Bishop, M. Steffen, J. M. Chow, and J. M. Gambetta, Rapid driven reset of a qubit readout resonator, Phys. Rev. Appl. **5**, 011001 (2016).

[59] A. Metelmann and A. Clerk, Nonreciprocal photon transmission and amplification via reservoir engineering, Phys. Rev. X **5**, 021025 (2015).

[60] L. Larger, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, and M. Jacquot, High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification, Phys. Rev. X **7**, 011015 (2017).

[61] L. Deng, The MNIST Database of handwritten digit images for machine learning research [Best of the Web], IEEE Signal Process. Mag. **29**, 141 (2012).