Practical trainable temporal post-processor for multi-state 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, NJ 08544, USA*

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

(Dated: October 31, 2023)

We develop and demonstrate a trainable temporal post-processor (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 only for 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, for signals subject to Gaussian white noise processes, the TPP provides a linearly-scaling semi-analytic generalization of matched filtering to an arbitrary number of states. The TPP can be efficiently, autonomously, and reliably trained on measurement data, and requires only linear operations, making it ideal for FPGA implementations in cQED for real-time processing of measurement data from general quantum systems.

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 multi-qubit 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 enhance 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 [9–14] and excitation to higher states [14, 15] 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 using schemes built for more standard readout paradigms [16]. 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 qubit state readout, standard post-processing of measurement records has remained relatively unchanged (with some exceptions [17, 18]): data is filtered using a "matched filter" (MF) constructed from the mean of measurement records for two states to be distinguished (for example, states $|e\rangle$ or $|g\rangle$ of a qubit). Crucially, the MF thus defined applies only to binary classification, and much more restrictively is optimal only if readout is subject to Gaussian white (i.e. uncorrelated) noise process [19]. In many cases, an even simpler (and less optimal) boxcar filter is employed, due to the ease of its construction. Our approach harnesses machine learning to provide a model-free trainable temporal post-processor (TPP) of quantum measurement data in general noise conditions, and for an arbitrary number of states of a generic measured quantum system ([20] for source code). We test our approach by applying it to the experimental readout of distinct qubits across a range of measurement powers. Our results show that the TPP reliably outperforms the standard MF under complex readout conditions at high powers, providing in certain cases a reduction in errors by a factor of several. Furthermore, the TPP achieves this improvement while requiring only linear weights applied to quantum measurement data (see Fig. 1): this makes it compatible with FPGA implementations for real-time hardware processing, and exacts a lower training cost than neural network-based machine learning schemes [21, 22].

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 [23], and chaotic systems' forecasting [24–26] and control [27]. 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 be-

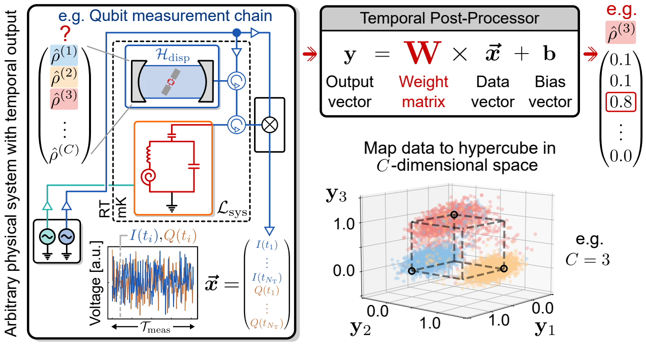


Figure 1. **Temporal post-processor (TPP) for multi-state classification using quantum measurement data, demonstrated for dispersive qubit readout in cQED.** The objective is to process temporal data corresponding to an unknown state (indexed σ) of an arbitrary physical system - here the state of a qubit 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 $\vec{x}$ 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 the TPP.

tween states of a multi-level 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 improve 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 'learnt' by the TPP from a given dataset. Secondly, TPP learning is tested on simulated quantum measurement datasets using 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 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 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 filters can deviate substantially from filters learned under the white noise assumption. Crucially, these noise-adapted TPP filters outperform generalized matched filters. By learning from quantum noise correlations, the TPP therefore utilizes a characteristic of quantum measurement data inaccessible to post-processing schemes relying on noise-agnostic matched filtering methods.

The established learning principles provide a structure to the general applicability of the TPP, which we believe 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, necessary on even industrial-grade quantum processors [28–30]. 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 quantum measurement task we use as an example to demonstrate the TPP: dispersive qubit readout in the cQED architecture. In Sec. III we then introduce our temporal post-processing framework to multi-state classification: a model-free supervised machine learning approach that can be applied to the classification of arbitrary time series. Importantly, we draw connections between the TPP approach and standard filtering-based approaches to qubit state measurement. In Sec. IV, we apply the developed TPP framework to experimental data for qubit readout, showing that it can outperform standard matched-filtering at strong measurement powers relevant for high-fidelity readout. Sec. V delves into the aspects that enable the TPP to learn filters that can be more effective than standard matched filters using controlled simulations. We conclude with a discussion on the general applicability of TPP for quantum state classification and temporal processing of quantum measurement data.

II. STANDARD POST-PROCESSING FOR DISPERSIVE QUBIT READOUT

A. Quantum measurement chain for dispersive qubit readout

The standard quantum measurement chain for heterodyne readout in cQED is depicted schematically in Fig. 1, and can be modeled via the stochastic master equation (SME):

$$d\hat{\rho}_c = \mathcal{L}_{\text{sys}}\hat{\rho}_c \, dt + \mathcal{L}_{\text{envt}}\hat{\rho}_c \, dt + \mathcal{L}_{\text{meas}}[dW]\hat{\rho}_c. \quad (1)$$

Here the Liouvillian superoperator $\mathcal{L}_{\text{sys}}$ defines the quantum system whose states are to be read out. We emphasize that the TPP approach enables classification for completely general $\mathcal{L}_{\rm sys}$. For relevance to cQED applications, in this paper we choose to focus on dispersive qubit readout, where the system comprises a multi-level artificial atom (here, a transmon) dispersively coupled to a readout cavity that is driven using a coherent tone at frequency ω_d. Then, $\mathcal{L}_{\rm sys}\hat{\rho} = -i[\hat{\mathcal{H}}_{\rm disp}, \hat{\rho}]$, where the dispersive Hamiltonian $\hat{\mathcal{H}}_{\rm disp}$ for a multi-level transmon takes the form (for cavity operators in the interaction frame with respect to ω_d and setting $\hbar = 1$)

$$\hat{\mathcal{H}}_{\rm disp} \simeq \sum_p \omega_p |p\rangle\langle p| - \Delta_{da}\hat{a}^\dagger\hat{a} + \sum_p \chi_p \hat{a}^\dagger\hat{a}|p\rangle\langle p|. \quad (2)$$

Here $\Delta_{da} = \omega_d - \omega_a$ is the detuning between the cavity and the readout tone at frequency ω_d, while χ_p is the dispersive shift per photon when the artificial atom is in state $|p\rangle$ [31, 32]. The general Liouvillian $\mathcal{L}_{\rm envt}$ is then used to describe all losses through channels that are not directly monitored, such as transmon transitions.

The final superoperator $\mathcal{L}_{\rm meas}$ 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 labelled $\hat{d}$. In the simplest case, $\mathcal{L}_{\rm meas}$ 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, $\mathcal{L}_{\rm meas}$ can describe the monitoring of several modes of a general quantum nonlinear processor that is embedded in the measurement chain [5]. Crucially, $\mathcal{L}_{\rm meas}$ 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:

$$I^{(\sigma)}(t_i) = \sqrt{\kappa}\langle\hat{X}^{(\sigma)}(t_i)\rangle_c + \xi_I(t_i) + \xi_I^{\rm cl}(t_i), \quad (3a)$$

$$Q^{(\sigma)}(t_i) = \sqrt{\kappa}\langle\hat{P}^{(\sigma)}(t_i)\rangle_c + \xi_Q(t_i) + \xi_Q^{\rm cl}(t_i). \quad (3b)$$

We consider discretized temporal indices t_i, for $i \in [N_{\rm T}]$ and $N_{\rm T} = \mathcal{T}_{\rm meas}/\Delta t$, where $\mathcal{T}_{\rm meas}$ is the total measurement time and Δt is the sampling time set by the digitizer. Heterodyne measurement probes the 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 mode $\hat{d}$ being monitored. More precisely, $\langle\hat{X}^{(\sigma)}(t_i)\rangle_c$, $\langle\hat{P}^{(\sigma)}(t_i)\rangle_c$ describe individual quantum trajectories of measured quadratures, conditioned on measurement records via a dependence on the heterodyne measurement noise $\xi_{I/Q}(t_i)$ through $\mathcal{L}_{\rm meas}$. The heterodyne measurement noise itself is modelled as zero-mean Gaussian white noise,

$$\mathbb{E}[\xi_{I,Q}(t_i)] = 0, \quad \mathbb{E}[\xi_{I,Q}(t_i)\xi_{I,Q}(t_j)] = \frac{1}{\Delta t}\delta_{ij}\delta_{I,Q} \quad (4)$$

where $\mathbb{E}[\cdot]$ describes ensemble averages over distinct noise realizations (obtained for distinct measurements). In contrast, the quantum trajectories contain the quantum noise contributions to the measurement records, in addition to state information: these include amplified quantum fluctuations when measuring the output field from a quantum amplifier, or the influence of quantum jumps in the measured cavity field due to transitions of the dispersively coupled qubit.

Finally, $\xi_{I/Q}^{\rm cl}(t_i)$ 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, they are formally distinct from heterodyne measurement noise as they are not associated with a stochastic measurement superoperator in Eq. (1).

The objective of the readout task is then to use this noisy temporal measurement data to obtain an estimated class label $\sigma^{\rm est}$ that is ideally equal to the true class label σ. Furthermore, we require *single-shot* readout [33], where the estimation must be performed using only a single measurement shot: such rapid readout is essential for quantum feedback and control applications [34–36].

B. Binary qubit state measurement and matched filters

The standard classification paradigm in cQED to obtain $\sigma^{\rm est}$ from raw heterodyne records would formally be described as a filtered Gaussian discriminant analysis (FGDA) in contemporary learning theory. This comprises two stages: (i) temporal filtering of each measured quadrature, and (ii) assigning a class label to filtered quadratures that maximises the likelihood of their observation amongst all C classes as determined by a Gaussian probability density function. Formally, this procedure can be written as:

$$\sigma^{\rm est} = {\rm G}\left[\sum_i \begin{pmatrix} h_I(t_i)I^{(\sigma)}(t_i) \\ h_Q(t_i)Q^{(\sigma)}(t_i) \end{pmatrix}\right] = {\rm G}\left[\begin{pmatrix} \vec{h}_I^T\vec{I}^{(\sigma)} \\ \vec{h}_Q^T\vec{Q}^{(\sigma)} \end{pmatrix}\right] \quad (5)$$

where in the second expression we have introduced vectorized notation in the space of measurement records, so that $\vec{I}_i = I(t_i)$, enabling the filtering step to be written as an inner product. The function ${\rm G}[\cdot]$ 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 using a calibration dataset: a set of $N_{\rm train}$ 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 [19, 37, 38]. The empirical matched filter is constructed from the calibration dataset, where (n) indexes distinct records, via

$$\vec{h}_I = \frac{1}{N_{\mathrm{train}}} \sum_{n=1}^{N_{\mathrm{train}}} \left(\vec{I}^{(e)}_{(n)} - \vec{I}^{(g)}_{(n)} \right) \tag{6}$$

with $\vec{h}_Q$ defined analogously with $I \to Q$. The function $\mathrm{G}[\cdot]$ requires fitting 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 [39]. Due to their ease of construction, often a matched filter akin to Eq. (6), or an even more rudimentary boxcar filter (a uniform filter that is nonzero only when the measurement signal is on) are deployed, regardless of the complexity of the noise conditions (for example, when qubit decay is significant and more optimal filters can be found [19]). We will show how the TPP approach provides a natural generalization of matched filtering to multi-state classification, and furnishes a trainable classifier that can generalize to more complex noise environments.

III. TRAINABLE TEMPORAL POST-PROCESSOR FOR MULTI-STATE CLASSIFICATION

Machine learning using only linear trainable weights has shown remarkable success in time-dependent supervised machine learning tasks [21]. In such cases, the objective is to map a time series faithfully to a dynamically-evolving target function via the application of an efficiently trainable *linear* transformation [22]. 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 [19].

To overview its key features we first introduce the mathematical framework underpinning the 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 $\vec{\boldsymbol{x}}^{(\sigma)} \in \mathbb{R}^{N_{\mathrm{O}} \cdot N_{\mathrm{T}}}$. As an example, in the case of heterodyne measurement, $N_{\mathrm{O}} = 2$ and $\vec{\boldsymbol{x}}^{(\sigma)} = \left(\begin{smallmatrix} \vec{I}^{(\sigma)} \\ \vec{Q}^{(\sigma)} \end{smallmatrix} \right)$ (see 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, $\mathbb{R}^{N_{\mathrm{O}} \cdot N_{\mathrm{T}}}$, 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 $\mathrm{F}[\cdot]$ on this vector $\mathbf{y}$, so that the complete transformation is:

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

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

$$\sigma^{\mathrm{est}} = \mathrm{F}[\mathbf{y}] = \mathrm{F}\Big[\mathbf{W}\vec{\boldsymbol{x}}^{(\sigma)} + \mathbf{b}\Big] \tag{7}$$

The function $\mathrm{F}[\cdot]$ 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 suitably-trained Gaussian discriminator $\mathrm{G}[\cdot]$ as in Eq. (5). The dimensions of the various components making up the TPP framework are summarized in Table I.

We note that at first sight Eq. (7), which defines the TPP scheme for classification, appears to be analogous to Eq. (5) in the FGDA scheme. There are, in fact, close connections between the two, as we will expand upon shortly. However, the TPP framework is also markedly different, in what can broadly be categorized as two aspects.

First, the defining feature of any machine learning approach: the ability (and requirement) to learn from data. $\mathbf{W} \in \mathbb{R}^{C \times N_{\mathrm{O}} \cdot N_{\mathrm{T}}}$ is a trainable matrix of weights and $\mathbf{b} \in \mathbb{R}^C$ is a vector of trainable biases, both learned from data $\vec{\boldsymbol{x}}^{(p)}$ with *known* labels p (C in total) in a supervised learning framework. More precisely, 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 $\mathbf{W}^{\mathrm{opt}}, \mathbf{b}^{\mathrm{opt}}$ minimize a least-squares cost function to achieve this target with minimal error:

$$\{\mathbf{W}^{\mathrm{opt}}, \mathbf{b}^{\mathrm{opt}}\} = \underset{\mathbf{W},\mathbf{b}}{\mathrm{argmin}} ||\mathbf{Y} - (\mathbf{W}\mathbf{X} + \mathbf{b})||^2 \tag{8}$$

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). The FGDA scheme using matched filters is in principle tailored to situations where useful signal in data is obscured only by additive Gaussian white noise, although it is applied much more broadly in practice. The TPP places no such restrictions on the training data *a priori*, and can therefore generalize to more nontrivial noise conditions, as we will show. Furthermore, a distinguishing feature of the TPP framework amongst other ML paradigms is that its optimization is convex and hence guaranteed to converge.

The second defining feature is the scope of applicability of the TPP framework. It natively generalizes to the classification of an arbitrary number of states C. Furthermore, no restriction is placed on the type of data that constitutes the vector $\vec{\boldsymbol{x}}$. In particular, no underlying physical model of the system generating the measurement data is *a priori* required: any relevant information must be learned by the TPP from data during the training phase. This also implies that the results in this paper apply to the classification of time series that have nothing to do with qubit state measurement. Its generality and ease of training enable the TPP to serve as a versatile trainable classifier, suited to a variety of classification tasks.

A. TPP learning mechanism and interpretation as optimal filtering

While Eq. (7) 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 this stochastic measurement data can be written in the very general form:

$$\vec{\boldsymbol{x}}^{(\sigma)} = \vec{\boldsymbol{s}}^{(\sigma)} + \vec{\boldsymbol{\zeta}}^{(\sigma)} \tag{9}$$

Here $\vec{\boldsymbol{\zeta}}^{(\sigma)}$ describes the stochasticity of the measured data: for heterodyne measurement, for example, this includes the noise sources from Eq. (3a), (3b), including quantum noise. We take the noise process to have zero mean, $\mathbb{E}[\vec{\zeta}_j^{(\sigma)}] = 0$. Then, $\vec{\boldsymbol{s}}^{(\sigma)} = \mathbb{E}[\vec{\boldsymbol{x}}^{(\sigma)}]$ are simply the mean traces of the measured data for state σ. Crucially, the noise is characterized by nontrivial second-order temporal correlations, which we define as $\boldsymbol{\Sigma}_{jk}^{(\sigma)} = \mathbb{E}[\vec{\zeta}_j^{(\sigma)} \vec{\zeta}_k^{(\sigma)}]$. Higher-order correlations of the noise will also be generally non-zero, but are not relevant for the discussion here.

The use of a least-squares cost function in Eq. (8) means that a closed form of the optimal weights $\mathbf{W}^{\rm opt}$ and biases $\mathbf{b}^{\rm opt}$ learned by the TPP can be obtained (see Appendix D). Furthermore, the form of Eq. (9) allows us to write these learned weights and biases as

$$\begin{pmatrix} \mathbf{W}^{\rm opt} & \mathbf{b}^{\rm opt} \end{pmatrix} = \mathbf{M}\mathbf{D}^{-1}. \tag{10}$$

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:

$$\mathbf{D} = \begin{pmatrix} \mathbf{G}+\mathbf{V} & \sum_c \vec{\boldsymbol{s}}^{(c)} \\ \sum_c (\vec{\boldsymbol{s}}^{(c)})^T & C \end{pmatrix} \tag{11}$$

which depends on the the "Gram" matrix of mean traces, $\mathbf{G} = \sum_c \vec{\boldsymbol{s}}^{(c)} (\vec{\boldsymbol{s}}^{(c)})^T$, but also on the temporal correlations via the matrix $\mathbf{V} = \sum_c \boldsymbol{\Sigma}^{(c)}$. Both these quantities emerge naturally in the analysis of the resolvable expressive capacity of noisy physical systems [40]. Here, Eq. (10) 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 $\mathbf{V}$. We will explore this dependence in the rest of our analysis.

Secondly, we find that the operation of TPP weights on data can be recast to clarify its connections to standard filtering-based classification schemes. To do so, we note that the learned matrix of weights $\mathbf{W}^{\rm opt} \in \mathbb{R}^{C \times N_{\rm O} \cdot N_{\rm T}}$ can be equivalently expressed as:

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

where $\vec{\boldsymbol{f}}_k \in \mathbb{R}^{N_{\rm O} \cdot N_{\rm T}}$ for $k \in [C]$. With this parameterization, Eq. (7) for the kth component of the vector $\mathbf{y}$ can be rewritten as:

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

When compared against Eq. (5), the interpretation of $\vec{\boldsymbol{f}}_k$ becomes clear: this set of weights can be viewed as a temporal filter applied to the data $\vec{\boldsymbol{x}}$. As a result, TPP based classification can equivalently be interpreted as the application of C filters (one for each k) to obtain the estimated label $\sigma^{\rm est}$. The optimal $\mathbf{W}^{\rm opt}$ 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. Furthermore, we note that the C filters are not all independent; they can be shown to satisfy the constraint (see Appendix D)

$$\sum_{k=1}^{C} \vec{\boldsymbol{f}}_k = \vec{\mathbf{0}}, \tag{14}$$

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

B. TPP-learned optimal filters for multi-state classification under Gaussian white noise

We begin by analyzing the case most often considered in cQED measurement chains, where the dominant noise source in heterodyne records I, Q is Gaussian white noise, which is assumed to be state and time-independent. Engineering of cQED measurement chains is geared towards approaching this limit, by (i) developing large bandwidth, high dynamic range amplifiers that operate with fast response times and minimal nonlinear effects even at high gain and large input signal powers [41, 42],[43–46], (ii) improving qubit T_1 and tolerance to strong cavity drives to reduce transitions during $\mathcal{T}_{\rm meas}$ [3], and (iii) controlling technical noise sources such as electronic white

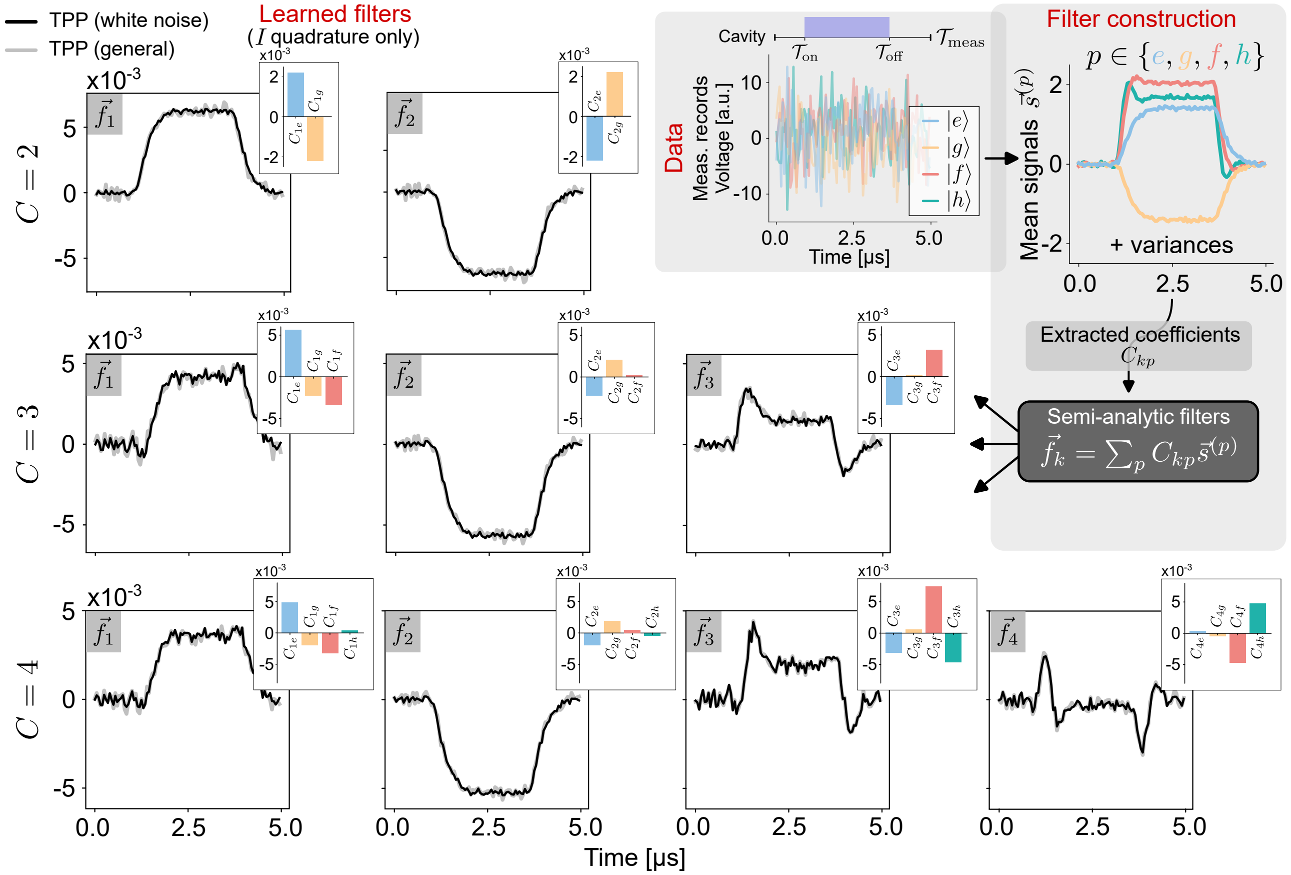


Figure 2. **TPP-learned optimal filters for simulated multi-state classification under Gaussian white noise conditions.** Top right: single-shot measurement records obtained under the indicated measurement tone, and empirical mean traces of several heterodyne records of the cavity I quadrature corresponding to multi-level 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 are filters learned under the white noise assumption, calculated analytically using Eq. (15). Bar plots show the coefficients C_{kp} applied to respective mean traces in calculating these filters. Gray curves are general filters calculated by numerically solving Eq. (8). Both analytically-computed white noise filters and general filters can be extended to arbitrary C.

noise from classical cryo-HEMT amplifiers and room temperature electronics.

In this relevant limit, we show that the C filters defined in Eq. (13) can be computed via:

$$\vec{\boldsymbol{f}}_k = \sum_{p\in\{e,g,\ldots\}} C_{kp}\vec{\boldsymbol{s}}^{(p)}, \quad k \in [C] \tag{15}$$

where $\vec{\boldsymbol{s}}^{(p)}$ are the empirically-calculated mean traces under the known initial state p:

$$\vec{\boldsymbol{s}}^{(p)} = \frac{1}{N_{\rm train}} \sum_{n=1}^{N_{\rm train}} \vec{\boldsymbol{x}}^{(p)}_{(n)} \tag{16}$$

while the coefficients C_{kp} can also be shown to depend only on $\vec{\boldsymbol{s}}^{(p)}$ and additionally the variance of the measured heterodyne records, assumed to be observable-independent and time-invariant, as mentioned earlier. Formally, here the correlation matrix $\mathbf{V}$ becomes proportional to the identity matrix (see Appendix D for full details). The TPP-learned optimal filters in the Gaussian white noise approximation are therefore simply a semi-analytically calculable linear combination of the mean traces. As a result, obtaining these optimal filters only requires the calculation of an empirical mean of the measurement records for each state, and an empirical estimate of the variances.

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. (1) to generate a dataset of measured heterodyne records for C qubit states, under the following assumptions: (i) all qubit state transitions are neglected, (ii) any additional classical noise sources in the measurement chain are ignored, and (iii) 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

$\mathcal{T}_{\rm meas}$, namely for $[\mathcal{T}_{\rm on}, \mathcal{T}_{\rm off}]$ (see Fig. 2, top right), and to be coincident with the cavity center frequency so that $\Delta_{da} = 0$, usual for transmon readout (for full details, see Appendix B 1). Other system parameters can be found in the caption of Fig. 2. We note that the specific details of the readout scheme do not change the TPP learning procedure. These simulations yield single-shot measurement records for any number of transmon states. Examples of these records are then shown in Fig. 2 for four distinct transmon states $p \in \{e, g, f, h\}$; for ease of visualization we only consider 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. (15). While the individual measurement records are obscured by white noise, the empirically-calculated mean traces in the top right of Fig. 2 illustrate the physics at play. The mean traces grow once the measurement tone is turned on past $\mathcal{T}_{\rm on}$, and settle to a steady state depending on the induced dispersive shift χ_p and the measurement amplitude. The traces begin to fall beyond $\mathcal{T}_{\rm off}$ 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 kth 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 shown in black in the top row of Fig. 2, together with bar plots showing the coefficients C_{kp}. Again for visualization, we only show filters $\vec{f}_k \in \mathbb{R}^{N_{\rm T}}$ for I quadrature data; the complete vector $\vec{\boldsymbol{f}}_k$ includes filters for all $N_{\rm O}$ 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 (see Appendix D). We note that the second filter ($k = 2$) is simply the negative of the first, as demanded by Eq. (14).

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 non-zero 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 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 is a substantially lower complexity relative to the pseudoinverse calculation demanded by Eq. (8), which requires inverting a much larger matrix in $\mathbb{R}^{N_{\rm O}\cdot N_{\rm T}\times N_{\rm O}\cdot N_{\rm T}}$ (see Appendix D).

Of course, the latter approach of obtaining $\mathbf{W}^{\rm opt}$ and hence TPP filters using Eq. (8) can also be employed for learning using the same training data. Here, it yields the underlying filters 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 TPP filters via Eq. (15), but is not *a priori* known to the general RC. 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 $\mathbf{W}^{\rm opt}$ accordingly. The fact that the general TPP filters approach 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 we will demonstrate in Secs. IV, V.

C. TPP performance under Gaussian white noise in comparison to standard FGDA

We now analyze the classification performance using TPP-learned optimal filters from the previous section in comparison to the standard FGDA approach. For concreteness, we perform dispersive qubit readout to distinguish $C = 3$ states $p \in \{e, g, f\}$. Recall that we consider the measurement tone to be resonant with the cavity, as is often the case for transmon readout. Then, the sign of cavity dispersive shifts for transmon states e and f is the same, and is opposite to that for g, making them harder to distinguish (see also Fig. 3 inset).

For this three-state classification task, a unique filter choice for the FGDA is not known. While certain approaches at constructing filters have been attempted [47], boxcar filtering is still commonly employed. Another approach might be to use a matched filter that optimizes distinction of just one pair of states. There are 3 such filters in total: for discrimination of e-g states as defined in Eq. (6), as well as analogously-defined filters for e-f and g-f states.

In Fig. 3, we show classification infidelities $1 - \mathcal{F}$, calculated for datasets with increasing measurement tone amplitude (more opaque markers), using both the optimal TPP filter and the FGDA with the four aforementioned filter choices. We clearly observe that the FGDA infidelities for most choices are worse than the RC. Interestingly, the poorest performer is not the boxcar filter; instead, it is the e-g filter, which would be optimal if we were only distinguishing $\{e, g\}$ states, that 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 substantially confuses e and f states that 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 as well as the e-g filter would. Due to the specific driving condi-

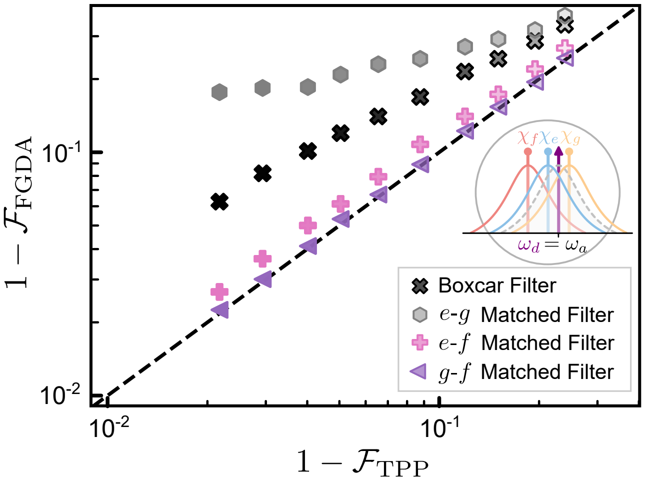


Figure 3. **Multi-state ($C = 3$) classification performance of TPP versus FGDA under Gaussian white noise conditions.** We consider dispersive qubit readout to distinguish states $p \in \{e, g, f\}$ as a function of measurement power; more opaque markers indicate stronger measurement tone amplitudes. Inset shows induced dispersive shifts for each state (not to scale). Standard FGDA is performed using 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, while all other chosen filters yield a worse performance.

tions 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 RC.

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. Furthermore, the number of distinct state pairs, and hence pairwise matched filters, grows quadratically with C in the absence of symmetries, making this brute force approach even less feasible for larger classification tasks. In contrast, the TPP approach provides a simple scheme to learn optimal filters that is automated, takes data for readout of all classes into account, and still scales 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 the conditions (i)-(iii) for qubit measurement chains listed in Sec. III B are not met. Departures from this ideal scenario are widely prevalent in cQED. Through 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

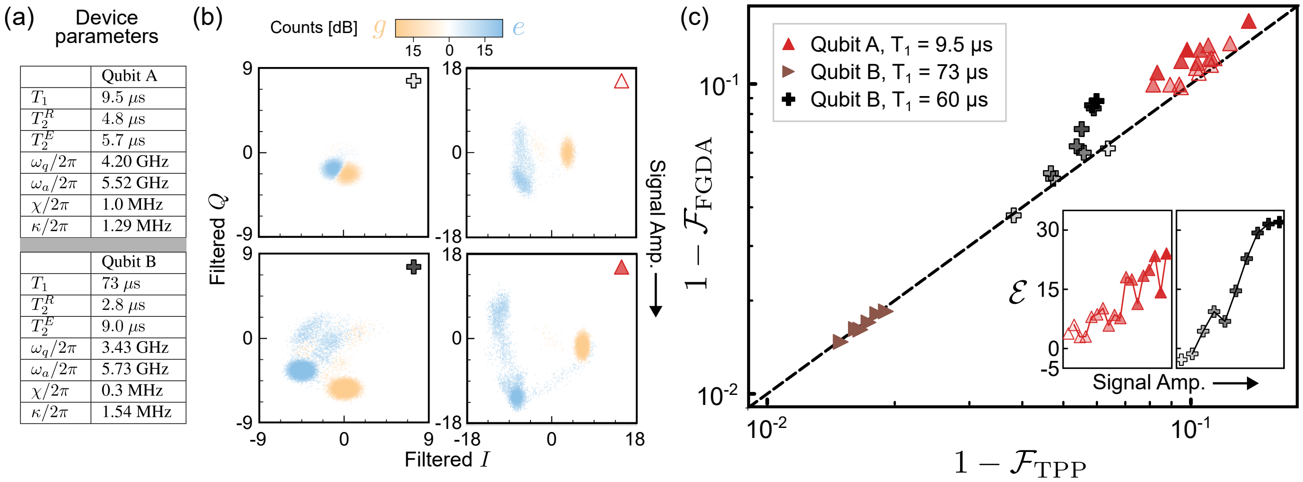
A. Experimental Results

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. (1). The actual circuit diagram is shown in Fig. 9 in Appendix A, and important parameters characterizing the measurement chain components are summarized in Fig. 4(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 $\mathcal{T}_{\mathrm{meas}}$, the ratio χ/κ determines the *theoretical* maximum readout fidelity; in particular, an optimal value for this ratio is known under these ideal conditions [32]. 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 $\bar{n}\kappa$ [48]. 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 [32]. 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. A8).

Each readout cavity is driven in reflection, and its output signal is amplified also in reflection using a Josephson Parametric Amplifier (JPA). We employ the latest iteration of strongly-pumped and weakly-nonlinear JPAs [46], boasting a superior dynamic range. Such JPAs operate well below saturation even at signal powers that correspond to over 100 photons, enabling us to probe qubit readout at high measurement powers. By choosing a signal frequency at 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 from half the pump frequency by greater 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 principle, the use of these stronger measurement tones should enhance the classification fidelity for qubit



	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

Figure 4. **Classification performance of 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 is analyzed and the resulting classification infidelities for binary ($C = 2$) state classification are plotted against each other. The dashed line marks $1 - \mathcal{F}_{\rm FGDA} = 1 - \mathcal{F}_{\rm TPP}$. For datasets with variable shading of markers (red and black), more opaque markers indicate stronger measurement tone amplitudes. Inset: Percentage fewer errors $\mathcal{E}$ computed for indicated datasets with increasing input signal amplitude.

readout. In practice, however, higher measurement powers are known to be associated with a variety of complex dynamical effects. 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 multi-level transitions to these higher levels can also be observed. There have also been predictions of chaotic dynamics and ionization [14, 49] at certain readout resonator occupation levels. The theoretical understanding of these effects, and their modeling via an SME analogous to Eq. (1) is an ongoing challenge.

In our experiments, we perform readout across this domain using measurement pulse durations ($\mathcal{T}_{\rm off} - \mathcal{T}_{\rm on}$) ranging from 500 ns to 900 ns, and measurement amplitudes from 0.04 to 0.09 in arbitrary voltage units, corresponding to roughly 44 to 100 photons in the cavity in the steady state. At the lowest pulse duration and 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 IQ 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. 4(a).

At the highest measurement powers, we are able to populate the readout cavity with up to 100 photons, calibrated by 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. 4(a) (see also Fig. 8 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. 4 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. (6)] under a variety of readout conditions, plotting the results against each other in Fig. 4.

The highest fidelity using 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 that any datasets above this line yield a higher classification *infidelity* with the FGDA than with the RC. 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 stronger the measurement power. For weaker measurement powers, we see that the TPP and the FGDA are once again comparable. However, a very clear trend emerges: for stronger measurement powers - where measurement dy-

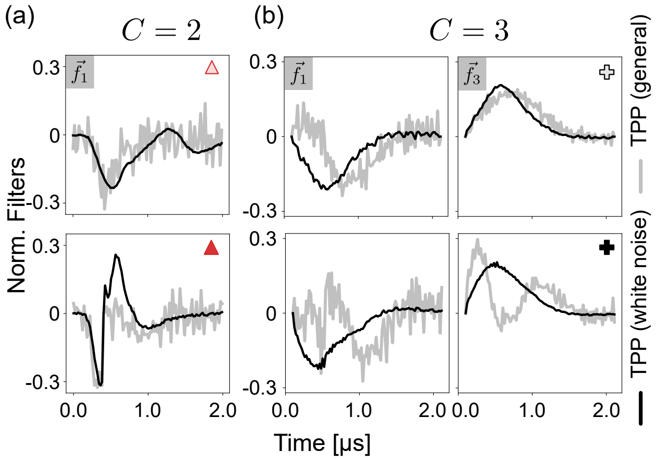


Figure 5. **Adaptation of TPP-learned filters with increasing measurement tone amplitude and evolving noise conditions.** Black curves are normalized TPP filters under the white noise assumption; for binary state classification, these are identical to standard matched filters. Gray curves are general TPP filters with no assumptions on noise statistics. (a) Filter $\vec{f}_1$ for binary ($C=2$) classification, and (b) filters $\vec{f}_{1,3}$ for $C=3$ state classification. In both cases, at weaker amplitudes the general TPP filter closely matches the TPP filter assuming white noise. However, for stronger measurement amplitudes, a marked difference between the white noise TPP filter and the general TPP filter is observed.

namics become much more complex as demonstrated in Fig. 4(a) - the TPP generally outperforms the FGDA.

To more precisely quantify the difference in performance between the TPP and FGDA, we introduce the metric $\mathcal{E}$:

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

which essentially asks: “what percentage fewer errors does the TPP make when compared to the FGDA?” We plot $\mathcal{E}$ in the inset of Fig. 4 for the two qubit readout experiments where the input signal amplitude is varied. We see clearly that with increasing amplitude, the TPP can significantly outperform the FGDA scheme, committing as many as 30% fewer errors in the experiments considered.

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 a useful tool in this vein.

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

The observed difference in performance between the TPP and the standard FGDA lies in the former's ability to learn from data as experimental conditions evolve. Our interpretation of TPP learning as the determination of optimal filters proves particularly insightful in expressing this adaptability.

Recall that for a C state classification task, the TPP learns C filters; however, the sum of filters is constrained by Eq. (14), so that $C-1$ filters are sufficient to describe the TPP's learning capabilities. In Fig. 5(a) we first consider filters learned by the TPP for a $C=2$ classification task, for select experimental datasets from Fig. 4 obtained under a low 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 are filters learned under the assumption of Gaussian white noise, given by Eq. (15); recall that for this binary case, these filters are exactly the standard MF. The gray curves, in contrast, are filters learned by the TPP for arbitrary noise conditions, obtained by solving Eq. (8). 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 earlier.

Crucially, the generalization of matched filters provided by TPP-learning via Eq. (15) enables a similar comparison for classification tasks of an arbitrary number of states. We show learned filters for $C=3$ state classification of $p \in \{e,g,f\}$ in Fig. 5(b), again for a low and 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.

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 strong measurement powers. However, the TPP is not provided with any physical description for such phenomena, which is in fact 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 in Eq. (10): 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.

V. TPP LEARNING: SIMULATION RESULTS

A. Learning correlations

As discussed in Sec. III A, the learned 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, when learning from data obtained from *quantum* systems, the observed correlations can 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: (i) 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 (ii) readout of finite lifetime qubits with multi-level transitions (quantum jumps).

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

Quantum-limited amplifiers are a mainstay of measurement chains in cQED, 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 anti-damping necessary for gain. For finite bandwidth amplifiers operating at large enough gains, this can lead to the addition of quantum noise with non-zero correlation time in measured heterodyne data.

To simulate qubit readout in these circumstances, we consider a quantum measurement chain described by Eq. (1) now consisting of a qubit-cavity-amplifier setup. $\mathcal{L}_{\text{meas}}$ then describes the readout of a non-degenerate (i.e. two-mode) parametric amplifier and its non-reciprocal coupling to the cavity used to monitor the qubit. We ignore qubit state transitions, so that $\mathcal{L}_{\text{envt}}$ only describes 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, $\xi^{\text{cl}}(t_i) = \sqrt{\bar{n}_{\text{cl}}}\frac{dW}{dt}(t_i)$, with a noise power $\bar{n}_{\text{cl}} = 30$, parameterized as usual in "photon number" units; these assumptions on the noise structure and power are taken from standard cQED experiments, including our own. Now, the obtained heterodyne measurement records, Eqs. (3a), (3b) contain two dominant noise sources: (i) excess classical white noise, and (ii) quantum noise added by the amplifier, contained once again in quantum trajectories $\langle \hat{X}^{(\sigma)}(t)\rangle_c$ and $\langle \hat{P}^{(\sigma)}(t)\rangle_c$.

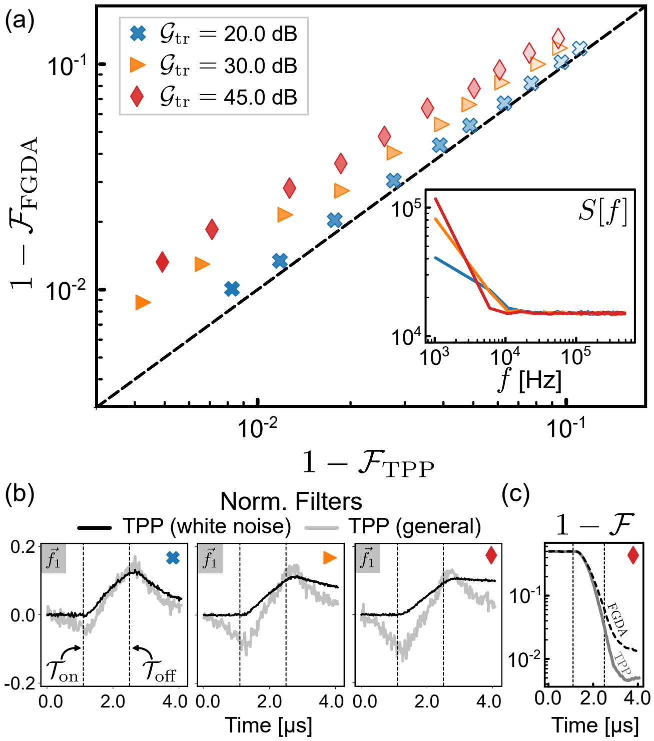


Figure 6. **Classification performance of TPP versus FGDA on simulated dataset of readout via a phase-preserving quantum amplifier.** (a) Classification infidelities for varying amplifier transmission gains $\mathcal{G}_{\text{tr}}$ as a function of measurement signal amplitude (more opaque markers are 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 white noise assumption (black) and general noise conditions (gray) for representative datasets of each value of $\mathcal{G}_{\text{tr}}$. (c) Classification infidelities as a function of total time t. The measurement tone is only on between the two vertical dashed lines.

We restrict ourselves for the moment to binary classification of states $|e\rangle$ and $|g\rangle$; here, the matched filtering (MF) scheme is unambiguously defined, and serves as a concrete benchmark for comparison to the TPP approach. In Fig. 6, we compare calculated infidelities using the FGDA and TPP approaches for three different values of amplifier transmission gain $\mathcal{G}_{\text{tr}}$, 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 I-quadrature) for state $|p\rangle$,

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

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 of Fig. 6; for the current readout task, this is independent of p. With increasing gain, the PSD deviates from the flat spectrum representative of white noise to a spectrum peaked at low frequencies, indicative of an extended correlation time. The observations also emphasize that added noise by the quantum amplifier dominates over heterodyne measurement noise ξ, as well as excess classical noise $\xi^{\rm cl}$.

For the lowest considered amplifier gain, we see that the FGDA and TPP classification performance is quite close. 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, shown in Fig. 6(b). The measurement tone is only on between the two dashed vertical lines. The curves in black show 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 non-zero even prior to the measurement signal turning on (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. 6(d) we plot $1-\mathcal{F}$ calculated for an increasing length of measured data, $t \in [0, \mathcal{T}_{\rm meas}]$. We clearly see that for $t < \mathcal{T}_{\rm on}$, both the TPP and FGDA cannot distinguish the states, as must be the case. The non-zero segment of the general TPP filter before $\mathcal{T}_{\rm on}$ instead accounts for noise correlations. In particular, due to the long correlation time of noise added by the quantum amplifier, noise in data beyond $\mathcal{T}_{\rm on}$ is correlated with noise from $t < \mathcal{T}_{\rm on}$. The general TPP filter is aware of these correlations that the standard MF is completely oblivious to, and by accounting for them improves classification performance.

C. Correlated quantum noise due to multi-level transitions

A transmon is a multi-level artificial atom, as described by Eq. (2); as a result, it is possible to excite levels beyond the typical two-level computational subspace of e and g states. Such transitions manifest 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 Fig. 7 inset). The transitions are described by superoperator $\mathcal{L}_{\rm envt}$, while $\mathcal{L}_{\rm meas}$ describes the measurement tone incident on the cavity, and the heterodyne measurement superoperator for the same; for full details see Appendix B 3.

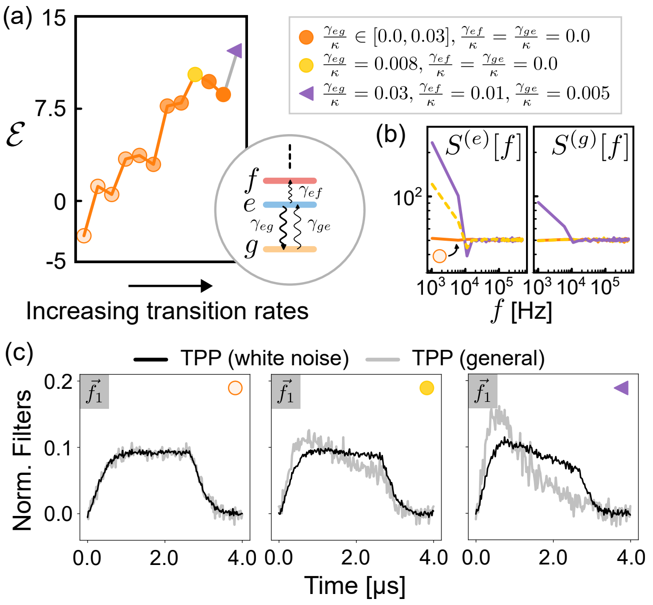


Figure 7. **Classification performance of TPP versus FGDA on simulated dataset of readout of a qubit experiencing multi-level transitions.** (a) $\mathcal{E}$ as a function of increasing transition rate values (more opaque markers), as shown. Schematic in the inset shows transmon levels and non-zero transition rates considered. (b) Noise PSD $S^{(p)}[f]$ for three representative datasets in the inset, indicating deviation from flat (white noise) as measurement data includes more transitions. (c) TPP learned filters (gray) compared to matched filters (black) for representative datasets, showing adaptation with transition rates.

For simplicity, we now further neglect excess classical noise added by the measurement chain, dropping terms $\xi^{\rm cl}_{I/Q}(t)$. As a result, the obtained measurement records, Eqs. (3a), (3b), 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. (1) 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. 7(a) with varying transition rates. The noise PSD is plotted in Fig. 7(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 only has 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 using the trained TPP in comparison to 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. 7(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 changed beyond means - in the correlations of measured data - and increasingly learns a distinct filter with sharply decaying features. We note that the utility of similar exponential linear filters for finite-lifetime qubits has been the subject of earlier analytic work [19]. The TPP approach generalizes the ability to learn such filters in the presence of arbitrary transition rates and measurement tones, and for multi-state classification.

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 reservoir computing 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 multi-level 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 RCs 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, 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 from practical systems where the majority 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 multi-level 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.

The TPP approach, anchored by its connection to standard matched filtering under simplified readout conditions, with demonstrated advantages for real qubit readout under more complex readout conditions, and feasible 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 learn from data could pave the way for an even broader class of applications. An important potential use is as a post-processor 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. The TPP is ideally suited to uncover the optimal post-processing step, through training that could be incorporated parallel to, or as part of, the optimization of the quantum machine. 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 would like to 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 and by the Army Research Office under Grant No. W911NF18-1-0144. 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, 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.

[1] A. Roy and M. Devoret, Introduction to parametric amplification of quantum signals with josephson circuits, Comptes Rendus Physique **17**, 740 (2016).

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

[3] A. P. M. Place, L. V. H. Rodgers, P. Mundada, B. M. Smitham, M. Fitzpatrick, Z. Leng, A. Premkumar, J. Bryon, A. Vrajitoarea, S. Sussman, G. Cheng, T. Madhavan, H. K. Babla, X. H. Le, Y. Gang, B. Jäck, A. Gyenis, N. Yao, R. J. Cava, N. P. de Leon, and A. A. Houck, New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds, Nature Communications **12**, 1779 (2021), number: 1 Publisher: Nature Publishing Group.

[4] G. Angelatos, S. A. Khan, and H. E. Türeci, Reservoir Computing Approach to Quantum State Measurement, Physical Review 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 [quant-ph] 10.48550/arXiv.2110.13849 (2021).

[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, Communications Physics **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, Physical Review Letters **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, Advanced Quantum Technologies **4**, 2100027 (2021).

[9] D. Sank, Z. Chen, M. Khezri, J. Kelly, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, T. White, J. Wenner, A. N. Korotkov, and J. M. Martinis, Measurement-induced state transitions in a superconducting qubit: Beyond the rotating wave approximation, Physical Review Letters **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, Physical Review B **101**, 134509 (2020), publisher: American Physical Society.

[11] A. Petrescu, M. Malekakhlagh, and H. E. Türeci, Lifetime renormalization of driven weakly anharmonic superconducting qubits. II. The readout problem, Physical Review B **101**, 134510 (2020).

[12] R. Hanai, A. McDonald, and A. Clerk, Intrinsic mechanisms for drive-dependent Purcell decay in superconducting quantum circuits, Physical Review Research **3**, 043228 (2021).

[13] M. Khezri, A. Opremcak, Z. Chen, A. Bengtsson, T. White, O. Naaman, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, A. L. Crook, B. Curtin, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, B. Foxen, G. Garcia, W. Giang, M. Giustina, R. Gosula, A. G. Dau, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, J. Kelly, S. Kim, P. V. Klimov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, M. McEwen, A. Megrant, X. Mi, K. C. Miao, S. Montazeri, A. Morvan, M. Neeley, C. Neill, A. Nersisyan, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Quintana, C. Rocque, P. Roushan, K. Sankaragomathi, K. J. Satzinger, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, N. Zhu, N. Zobrist, D. Sank, A. Korotkov, Y. Chen, and V. Smelyanskiy, Measurement-induced state transitions in a superconducting qubit: Within the rotating wave approximation, arXiv:2212.05097 [quant-ph] (2022).

[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, Physical Review Applied **18**, 034031 (2022).

[15] 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, Physical Review Applied **15**, 064030 (2021).

[16] 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, Physical Review Applied **7**, 054020 (2017), publisher: American Physical Society.

[17] M. Tsang, Volterra filters for quantum estimation and detection, Physical Review A **92**, 062119 (2015).

[18] 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, Physical Review Applied **17**, 014024 (2022), publisher: American Physical Society.

[19] 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, Physical Review A **76**, 012325 (2007), publisher: American Physical Society.

[20] Source code available at `https://zenodo.org/doi/10.5281/zenodo.10020462`.

[21] 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 Networks **115**, 100 (2019).

[22] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa, Next generation reservoir computing, Nature Communications **12**, 5564 (2021), number: 1 Publisher: Nature Publishing Group.

[23] D. Canaday, A. Griffith, and D. J. Gauthier, Rapid time series prediction with a hardware-based reservoir computer, Chaos: An Interdisciplinary Journal of Nonlinear Science **28**, 123119 (2018), publisher: American Institute of Physics.

[24] 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: An Interdisciplinary Journal of Nonlinear Science **27**, 121102 (2017).

[25] 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, Physical Review Letters **120**, 024102 (2018).

[26] A. Griffith, A. Pomerance, and D. J. Gauthier, Forecasting chaotic systems with very low connectivity reservoir computers, Chaos: An Interdisciplinary Journal of Nonlinear Science **29**, 123108 (2019).

[27] D. Canaday, A. Pomerance, and D. J. Gauthier, Model-free control of dynamical systems with deep reservoir computing, Journal of Physics: Complexity **2**, 035025 (2021), publisher: IOP Publishing.

[28] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, Physical Review A **93**, 060302 (2016).

[29] 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 [quant-ph] (2018).

[30] 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).

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

[32] A. Blais, A. L. Grimsmo, S. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Reviews of Modern Physics **93**, 025005 (2021).

[33] 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, Nature Physics **5**, 791 (2009), number: 11 Publisher: Nature Publishing Group.

[34] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, Initialization by Measurement of a Superconducting Quantum Bit Circuit, Physical Review Letters **109**, 050507 (2012).

[35] J. E. Johnson, C. Macklin, D. H. Slichter, R. Vijay, E. B. Weingarten, J. Clarke, and I. Siddiqi, Heralded State Preparation in a Superconducting Qubit, Physical Review Letters **109**, 050506 (2012).

[36] P. Campagne-Ibarcq, E. Flurin, N. Roch, D. Darson, P. Morfin, M. Mirrahimi, M. H. Devoret, F. Mallet, and B. Huard, Persistent Control of a Superconducting Qubit by Stroboscopic Measurement Feedback, Physical Review X **3**, 021008 (2013).

[37] G. Turin, An introduction to matched filters, IRE Transactions on Information Theory **6**, 311 (1960), conference Name: IRE Transactions on Information Theory.

[38] 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, Physical Review A **93**, 062310 (2016).

[39] 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 2018 558:7709 **558**, 264 (2018).

[40] 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, arXiv:2307.16083 [quant-ph] 10.48550/arXiv.2307.16083 (2023).

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

[42] 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, Physical Review Applied **8**, 054030 (2017), publisher: American Physical Society.

[43] 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, Physical Review Applied **17**, 034064 (2022).

[44] 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, Physical Review Applied **20**, 034027 (2023).

[45] R. Kaufman, T. White, M. I. Dykman, A. Iorio, G. Stirling, 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, arXiv:2305.17816 [quant-ph] https://doi.org/10.48550/arXiv.2305.17816 (2023).

[46] R. Kaufman, C. Liu, K. Cicak, B. Mesits, M. Xia, C. Zhou, M. Nowicki, D. Pekker, J. Aumentado, and M. Hatridge, In Preparation (2024).

[47] 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 Information 2023 9:1 **9**, 1 (2023).

[48] D. I. Schuster, A. Wallraff, A. Blais, L. Frunzio, R. S. Huang, J. Majer, S. M. Girvin, and R. J. Schoelkopf, Ac stark shift and dephasing of a superconducting qubit strongly coupled to a cavity field, Physical Review Letters **94**, 123602 (2005).

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

[50] A. Metelmann and A. A. Clerk, Nonreciprocal Pho-

ton Transmission and Amplification via Reservoir Engineering, arXiv:1502.07274 [cond-mat, physics:quant-ph] (2015), arXiv: 1502.07274.

[51] 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, Physical Review X **7**, 011015 (2017), publisher: American Physical Society.

APPENDICES

Appendix A: Experimental Setup

In this appendix section, we show a few more examples of readout IQ histograms as well as a more detailed circuit diagram for the measurement chain. Shown in Fig. 8 below, we see two examples of the extremes of the measurement data used to generate Fig.4. Part (a) shows a lower power readout pulse performed for a short 300ns time, 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 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 three 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 large 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.

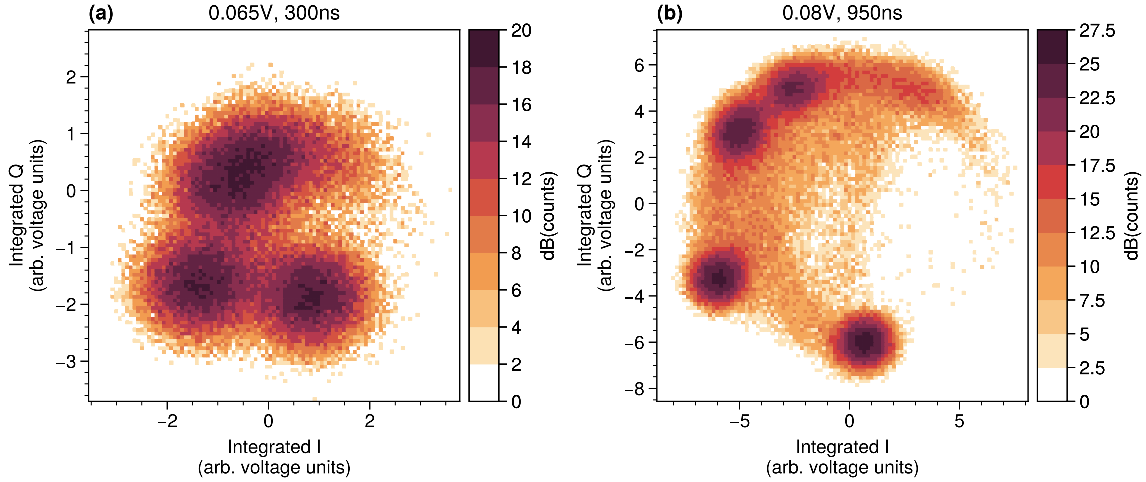


Figure 8. Comparison between boxcar-integrated IQ results of (a) a lower power pulse applied for a short time and (b) a higher power pulse applied for a longer time. State transitions are visible as "trails" leading between the primary symbols in (b). Counts are shown in logarithmic units to emphasize low-count trails.

In Fig. A9, the hardware schematic of the measurements in section IV are shown. The measurement setup is fairly standard, using single sideband upconversion to send signals into the dilution refrigerator, moving through three stages of attenuation with 20dB attenuation at 4K, 20dB attenuation at the 100mK stage, and approximately 45dB of attenuation at the base stage of the refrigerator, with 10dB 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, amplified in reflection, and then is routed once again back through the circulators to the remaining stages of amplification at 4K and room temperature accordingly. From there it is downconverted by the same local oscillator to 50MHz, filtered, amplified once more at low frequency, digitized at 1GS/s, and finally demodulated and integrated to acquire a readout histogram such as the ones shown in Fig. 8.

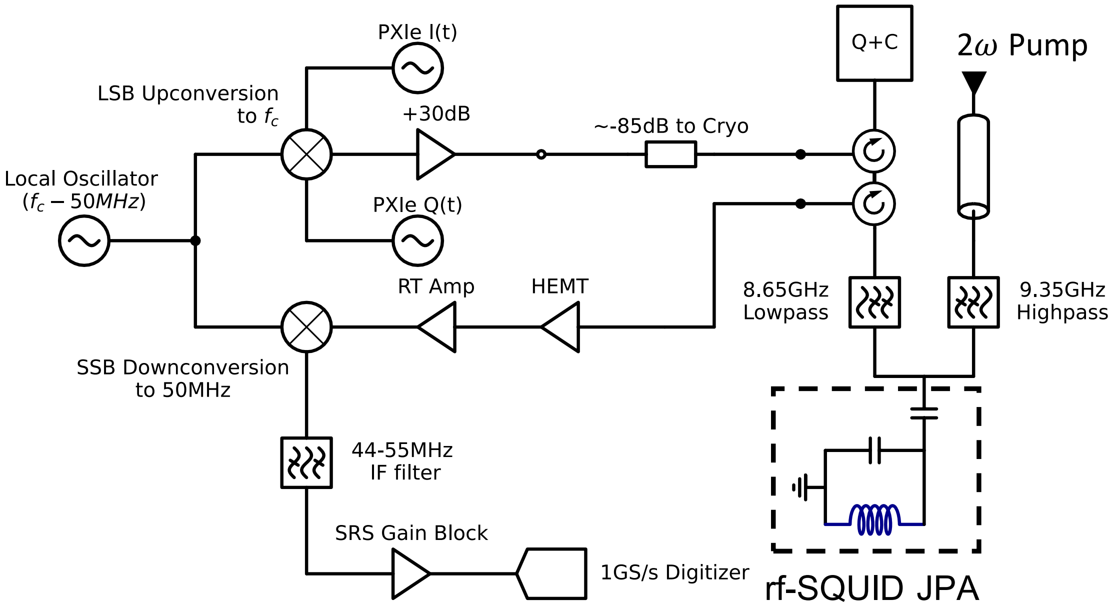


Figure 9. Upconversion and downconversion schematic for drive pulses sent first to the qubit readout resonator, driven in reflection, then routed to the amplifier and to the HEMT via two circulators.

Appendix B: Simulating heterodyne measurement records obtained from quantum measurement chains for dispersive qubit readout

In this appendix section, 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. (1):

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

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 multi-level 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 the 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.

Then, $\mathcal{S}[dW]$ 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 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 readout 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:

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

and by conjugation,

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

following which we arrive at:

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

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

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

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

$${\rm tr}_Q\{\mathcal{L}_{\rm q}(|p\rangle\langle p| \otimes \hat{\rho}_c)\} = {\rm tr}_Q\{|p\rangle\langle p| \otimes \mathcal{L}_{\rm q}\hat{\varrho}_c\} = {\rm tr}_Q\{|p\rangle\langle p|\} \otimes \mathcal{L}_{\rm q}\hat{\varrho}_c = \mathcal{L}_{\rm q}\hat{\varrho}_c \tag{B13}$$

As $\mathcal{L}_{\rm q}$ 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 tracing out the qubit sector yields:

$$\sqrt{\kappa}\ {\rm tr}_Q\{\mathcal{S}[dW]|p\rangle\langle p| \otimes \hat{\rho}_c\} = \sqrt{\kappa}\ {\rm tr}_Q\{|p\rangle\langle p|\} \otimes \mathcal{S}[dW]\hat{\varrho}_c = \sqrt{\kappa}\mathcal{S}[dW]\hat{\varrho}_c \tag{B14}$$

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

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

The resulting SME is purely linear and can be solved exactly using a truncated equations of motion (TEOMs) 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, $\mathcal{L}_{\rm envt} \to 0$, and the SME of Eq. (B1) takes the simpler form:

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

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

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

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

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

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

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

The superoperator $\mathcal{L}_c$ then defines the non-reciprocal coupling between the cavity mode and amplifier's signal mode $\hat{d}$,

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

To ensure non-reciprocal 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 $g = \Gamma$ [50].

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 $\kappa = \kappa' + \Gamma_c$. Then, we choose cavity parameters so that $\kappa' = \Gamma_c = 0.5\kappa$, and the dispersive shift $\chi/\kappa = 0.5$. Recall that perfect non-reciprocal 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 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 $\mathcal{L}_{\rm meas}$ 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:

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

for $\mathcal{L}_{\rm meas}$ now given by Eq. (B17). The resulting SME is again linear and can be solved exactly using a truncated equations of motion (TEOMs) approach.

3. Dispersive readout including multi-level transitions using a cavity

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

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

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

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

where γ_{jk} is the rate of transition from qubit state $|j\rangle$ to 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 can not generally be expressed in the form of Eq. (B8). Hence Eq. (B23) is integrated in joint the qubit-cavity Hilbert to generated 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. We begin with the TPP map defined in the main text, Eq. (7):

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

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: $\vec{x}_{(n)} \in \mathbb{R}^{N_{\text{O}} \cdot N_{\text{T}}}$ 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 $\vec{x}_1 = \vec{I}$, $\vec{x}_2 = \vec{Q}$. 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 carried out via two operations. First, the measurement records $\vec{\boldsymbol{x}}^{(\sigma)}_{(n)}$ are mapped to an intermediate target vector $\mathbf{y}^{(\sigma)}_{(n)}$ employing a 'one-hot' encoding (conventional for classification tasks): The kth element of this target vector $\mathbf{y}^{(\sigma)}_{(n)}$ is given by:

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

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 $\mathbf{X} \in \mathbb{R}^{N_{\text{O}} \cdot N_{\text{T}} \times C N_{\text{train}}}$:

$$\mathbf{X} = \begin{pmatrix} \vec{\boldsymbol{x}}^{(1)}_{(1)} & \vec{\boldsymbol{x}}^{(1)}_{(2)} & \cdots & \vec{\boldsymbol{x}}^{(1)}_{(N_{\text{train}})} & \cdots & \vec{\boldsymbol{x}}^{(C)}_{(1)} & \vec{\boldsymbol{x}}^{(C)}_{(2)} & \cdots & \vec{\boldsymbol{x}}^{(C)}_{(N_{\text{train}})} \end{pmatrix} \tag{C4}$$

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

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}^{(1)}_{(1)} & \mathbf{y}^{(1)}_{(2)} & \cdots & \mathbf{y}^{(1)}_{(N_{\text{train}})} & \cdots & \mathbf{y}^{(C)}_{(1)} & \mathbf{y}^{(C)}_{(2)} & \cdots & \mathbf{y}^{(C)}_{(N_{\text{train}})} \end{pmatrix} \tag{C5}$$

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

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}\vec{1} = \begin{pmatrix} \mathbf{W} & \mathbf{b} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \vec{1} \end{pmatrix} \equiv \boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{X}} \tag{C6}$$

Eq. (C6) helps us define $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{(N_{\text{O}} \cdot N_{\text{T}} + 1) \times C N_{\text{train}}}$ as a matrix which contains all measured records as well as a row of ones to account for the contribution of biases. Then, $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{C \times (N_{\text{O}} \cdot N_{\text{T}} + 1)}$ is the composite matrix of all learned weights. Eq. (C6) defines a regression problem that can be solved to obtain the optimal weights [51],

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{Y}\boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T)^{-1} \tag{C7}$$

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

$$N_{\text{train}}\mathbf{M} \equiv \mathbf{Y}\boldsymbol{\mathcal{X}}^T, \tag{C8}$$

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

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

so that Eq. (C7) can equivalently be written as:

$$\boldsymbol{\mathcal{W}}^{\mathrm{opt}} = \mathbf{M}\mathbf{C}^{-1} \tag{C10}$$

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 $\mathbf{C}^{+}$, related to the pseudoinverse of $\boldsymbol{\mathcal{X}}$, and defined by the following limit relation defining the pseudoinverse:

$$\mathbf{C}^{+} = \lim_{\lambda \to 0} (\mathbf{C} - \lambda \mathbf{I})^{-1} \tag{C11}$$

where $\mathbf{I}$ is the identity matrix on $\mathbb{R}^{(N_{\mathrm{O}} \cdot N_{\mathrm{T}}+1)\times(N_{\mathrm{O}} \cdot N_{\mathrm{T}}+1)}$ and λ is typically referred to as a regularization parameter. If $\mathbf{C}$ is invertible, we have $\mathbf{C}^{+} \to \mathbf{C}^{-1}$. We do emphasize that for the datasets analyzed in this paper, we consider a weak regularization of $\lambda = 10^{-15}$, if any.

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 $N_{\mathrm{train}} < N_{\mathrm{traj}}$ records as described above. The remaining $N_{\mathrm{test}} = N_{\mathrm{traj}} - N_{\mathrm{train}}$ records are used to construct a testing set. Predicted state labels are obtained using this testing set via both the FGDA scheme, Eq. (5) of the main text, and the TPP, Eq. (7). This process is repeated a total of $L = 10$ times: each time, a new set of weights $\boldsymbol{\mathcal{W}}^{\mathrm{opt}}$ is obtained from a distinct randomly chosen training set of the total N_{traj} records, and classification infidelities computed using the new random testing datasets. All classification fidelities are averaged to obtain the final values plotted in the main text. This approach is standard in machine learning, and ensures that the observed performance is not unduly effected by variations due to the specific training or testing dataset used.

Appendix D: TPP learned weights as optimal filters: analytic results

In this appendix, we will attempt to find an explicit form for the matrix $\boldsymbol{\mathcal{W}}^{\text{opt}}$ from the previous seciton, 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:

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

where $\vec{\boldsymbol{\zeta}}^{(c)}$ 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 $\vec{\xi}_{\text{cl}}$, as well as quantum noise in conditional quantum trajectories. Without loss of generality, $\vec{\boldsymbol{\zeta}}$ can always be taken to have zero mean,

$$\mathbb{E}[\vec{\boldsymbol{\zeta}}_j] = 0 \;\forall\; j \tag{D2}$$

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

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

The noise process will in general also possess non-zero 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 only have access to a finite number of shots N_{train}. Then, the above mean can be approximated using the estimator:

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

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

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

Assuming the very general form of Eq. (D1), we can proceed to greatly simplify the matrices $\mathbf{M}$ and $\mathbf{C}$.

a. *Simplification of mean matrix* $\mathbf{M}$

The mean matrix $\mathbf{M}$, Eq. (C8), can be written explicitly as

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

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

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

Here, we have used the fact that the sum over the columns of $\mathbf{Y}$ (and of $\mathbf{X}$), indexed by k, can be decomposed into two sums: over N_{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,

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

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

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

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 $\mathbf{C}$ is more involved. We begin by expanding it to the form:

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

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:

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

Next, we consider the off-diagonal block term,

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

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

Finally, we consider the block matrix,

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

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

$$[\mathbf{X}\mathbf{X}^T]_{ij} = \sum_c\sum_n[\vec{\boldsymbol{x}}^{(c)}_{(n)}]_i[\vec{\boldsymbol{x}}^{(c)}_{(n)}]_j = \sum_c\left\{[\vec{\boldsymbol{s}}^{(c)}]_i[\vec{\boldsymbol{s}}^{(c)}]_j + \sum_n[\vec{\boldsymbol{\zeta}}^{(c)}_{(n)}]_i[\vec{\boldsymbol{s}}^{(c)}]_j + [\vec{\boldsymbol{s}}^{(c)}]_i\sum_n[\vec{\boldsymbol{\zeta}}^{l(c)}_{(n)}]_j + \sum_n[\vec{\boldsymbol{\zeta}}^{(c)}_{(n)}]_i[\vec{\boldsymbol{\zeta}}^{(c)}_{(n)}]_j\right\} \tag{D15}$$

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

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

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 *covariance* matrix $\mathbf{V}$:

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

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

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

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 $\boldsymbol{\mathcal{W}}^{\text{opt}}$ 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:

$$\boldsymbol{\mathcal{W}}^{\text{opt}} \equiv \begin{pmatrix} (\vec{\boldsymbol{f}}^1)^T & b^1 \\ \vdots & \vdots \\ (\vec{\boldsymbol{f}}^C)^T & b^C \end{pmatrix} \equiv \begin{pmatrix} (\vec{F}_1)^T \\ \vdots \\ (\vec{F}_C)^T \end{pmatrix} \tag{D20}$$

Next, using Eq. (C7) together with the explicit form of the mean matrix $\mathbf{M}$ in Eq. (D10), we arrive at the important relation:

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

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 $\vec{\mathbf{0}} \in \mathbb{R}^{N_{\text{O}} \cdot N_{\text{T}}}$ is a vector of zeros. In particular, we wish to evaluate $\mathbf{C}\vec{n}$. Using the block representation of $\mathbf{C}$, we have:

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

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

Now, using Eq. (D22), multiplying through by $\mathbf{C}^{-1}$ 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 under the stationary quadrature-independent Gaussian white noise approximation: "matched filters" for arbitrary C

We find that under specific assumptions on the noise in measured data, the optimal learned filters by the TPP can be determined analytically using the strategy proposed in the previous section. The special case we find is one where $\vec{\boldsymbol{x}}^{(c)}$ in Eq. (D1) experience additive stationary Gaussian white noise. This noise process possesses a δ-function autocorrelation that further does not vary in time. We additionally assume the autocorrelation is identical for all $N_{\rm O}$ measured observables. The noise covariance properties of this process are then compactly given by:

$$\boldsymbol{\Sigma}^{(c)}_{jk} = \mathbb{E}[\vec{\boldsymbol{\zeta}}^{(c)}_j \vec{\boldsymbol{\zeta}}^{(c)}_k] = \Sigma^{(c)}\delta_{jk} \tag{D28a}$$

where $\Sigma^{(c)}$ is simply the stationary, observable-independent variance of the Gaussian white noise process. This naturally simplifies the form of the empirical noise covariance matrix $\mathbf{V}$ in Eq. (D18),

$$\mathbf{V} = \sum_c \Sigma^{(c)}\bar{\mathbf{I}} \equiv V\bar{\mathbf{I}} \tag{D29}$$

where $\bar{\mathbf{I}}$ is the identity matrix on $\mathbb{R}^{N_{\rm O}\cdot N_{\rm T}\times N_{\rm O}\cdot N_{\rm T}}$. Crucially, this noise process simplifies the second-order correlation matrix $\mathbf{C}$:

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

In particular, the contribution from noise correlations now appears as an identity matrix.

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 will once again make use of the simplified block representation of $\mathbf{C}$, which allows us to write:

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

It proves useful to define the overlap of mean traces

$$O_{cc'} = (\vec{\boldsymbol{s}}^{(c')})^T\vec{\boldsymbol{s}}^{(c)}, \tag{D32}$$

following which we can write:

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

Finally, defining

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

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

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

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, and multiply through by the regularized inverse of $\mathbf{C}$. This yields

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

Note that this approach foregoes the calculation of the regularized inverse of $\mathbf{C}$ in the computation of the learned filters $\vec{F}_c$. We emphasize here that if we instead consider observable-*dependent* Gaussian white noise, the terms $M_{cc'}$ are replaced by a block diagonal matrix in $\mathbb{R}^{(N_\mathrm{O}\cdot N_\mathrm{T}+1)\times(N_\mathrm{O}\cdot N_\mathrm{T}+1)}$. These matrices will not generally commute with $(\mathbf{C} - \lambda\mathbf{I})^{-1}$, preventing the transition from the first to the second line above, which is crucial to introducing $\vec{F}_c$ to the system. Filters for general situations such as that can always be obtained by evaluating Eq. (C7) numerically.

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 considering the difference of Eq. (D36) for any two distinct c values. For example, considering $c \neq c'' \in [1, \ldots, C]$:

$$\vec{S}^{(c)} - \vec{S}^{(c'')} = \sum_{c'} M_{cc'}\vec{F}_{c'} - \sum_{c'} M_{c''c'}\vec{F}_{c'} = \sum_{c'} [M_{cc'} - M_{c''c'}]\,\vec{F}_{c'} \tag{D37}$$

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 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 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 pth 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:

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

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

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

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 $\vec{F}_c$, we can immediately read off the individual filters for each measured quadrature:

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

The bias terms are finally given by:

$$\mathbf{b}_c = - \sum_p \mathbf{Q}^{-1}_{cp} \mathbf{T}_{pp} \tag{D46}$$

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

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

$$\vec{\boldsymbol{f}}_c = \mathbf{Q}^{-1}_{c1} \vec{\boldsymbol{s}}^{(1)} + \sum_{p=2}^{C-1} \left[\mathbf{Q}^{-1}_{cp} - \mathbf{Q}^{-1}_{c(p-1)} \right] \vec{\boldsymbol{s}}^{(p)} - \mathbf{Q}^{-1}_{c(C-1)} \vec{\boldsymbol{s}}^{(C)} \tag{D47}$$

which provides the learned filters as a linear combination of mean signals corresponding to each state to be classified. Comparing with Eq. (15) from the main text, we have:

$$\vec{\boldsymbol{f}}_c = \sum_{c=1}^{C} C_{cp} \vec{\boldsymbol{s}}^{(p)}, \ C_{cp} = \begin{cases} +\mathbf{Q}^{-1}_{c1} & \text{if } c = 1, \\ -\mathbf{Q}^{-1}_{c(C-1)} & \text{if } c = C, \\ \mathbf{Q}^{-1}_{cp} - \mathbf{Q}^{-1}_{c(p-1)} & \text{otherwise.} \end{cases} \tag{D48}$$

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:

$$\vec{S}^{(1)} - \vec{S}^{(2)} = [M_{11} - M_{21} - (M_{12} - M_{22})]\, \vec{F}_1 + [M_{21} - M_{22}]\, \vec{n} \tag{D49}$$

Hence we can directly read off:

$$\vec{F}_1 = \begin{pmatrix} \vec{\boldsymbol{f}}_1 \\ \mathbf{b}_1 \end{pmatrix} = \frac{1}{M_{11} - M_{21} - (M_{12} - M_{22})} \begin{pmatrix} \vec{\boldsymbol{s}}^{(1)} - \vec{\boldsymbol{s}}^{(2)} \\ 0 \end{pmatrix} - \frac{M_{21} - M_{22}}{M_{11} - M_{21} - (M_{12} - M_{22})} \begin{pmatrix} \vec{\mathbf{0}} \\ 1 \end{pmatrix} \tag{D50}$$

Appendix E: Time-shuffled data

As discussed in the previous section, the trained weights $\mathbf{W}$ take the form of Eq. (C7),

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{Y}\boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T - \lambda\mathbf{I})^{-1} \tag{E1}$$

We now consider the operation of a matrix $\mathbf{J}$ on $\mathbf{X}$ that serves to re-order 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 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{(N_{\text{O}}\cdot N_{\text{T}}+1)\times CN_{\text{train}}}$ and the exchange matrix is intended to switch *rows* of the data, $\mathbf{J} \in \mathbb{R}^{(N_{\text{O}}\cdot N_{\text{T}}+1)\times(N_{\text{O}}\cdot N_{\text{T}}+1)}$. Furthermore, the exchange matrix satisfies the properties: $\mathbf{J}^{-1} = \mathbf{J} = \mathbf{J}^T$, so that $\mathbf{J}\mathbf{J} = \mathbf{I}$.

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

$$\boldsymbol{\mathcal{X}}_J = \mathbf{J}\boldsymbol{\mathcal{X}} \implies \boldsymbol{\mathcal{X}} = \mathbf{J}\boldsymbol{\mathcal{X}}_J \tag{E2}$$

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.

Then, the trained weights can equivalently be written as:

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{Y}(\mathbf{J}\boldsymbol{\mathcal{X}}_J)^T(\mathbf{J}\boldsymbol{\mathcal{X}}_J\boldsymbol{\mathcal{X}}_J^T\mathbf{J}^T - \lambda\mathbf{I})^{-1} \tag{E3}$$

which, after some simplification and using $\mathbf{J}^T = \mathbf{J}$ reduces to:

$$\boldsymbol{\mathcal{W}}^{\text{opt}} = \mathbf{Y}\boldsymbol{\mathcal{X}}_J^T\mathbf{J}\mathbf{J}(\boldsymbol{\mathcal{X}}_J\boldsymbol{\mathcal{X}}_J^T)^{-1}\mathbf{J} = \left[\mathbf{Y}\boldsymbol{\mathcal{X}}_J^T(\boldsymbol{\mathcal{X}}_J\boldsymbol{\mathcal{X}}_J^T - \lambda\mathbf{I})^{-1}\right]\mathbf{J} \tag{E4}$$

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

$$(\boldsymbol{\mathcal{W}}^{\text{opt}})_J = \boldsymbol{\mathcal{W}}^{\text{opt}}\mathbf{J} \tag{E5}$$

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

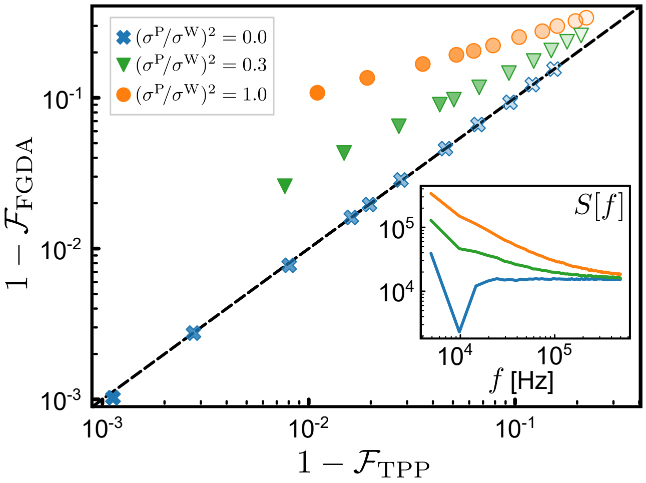


Figure 10. **Comparative classification performance of FGDA versus TPP in the presence of classical correlated noise.** We consider a $C = 2$ (binary) dispersive qubit readout task using simulated data and for different colored noise conditions. Darker markers indicate stronger measurement tone amplitudes. The dashed line indicates $1 - \mathcal{F}_{\mathrm{FGDA}} = 1 - \mathcal{F}_{\mathrm{TPP}}$. The inset plots the corresponding noise spectral density $S[f]$, which remains unchanged with coherent input power.

Appendix F: TPP learning of correlated *classical* noise

In this appendix 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. Like Sec. V B, we again consider simulated datasets of measured heterodyne records from a measurement chain of a qubit-cavity-amplifier setup, as in Sec. III B. 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):

$$\xi^{\mathrm{cl}}(t_i) = \sigma^{\mathrm{W}}\xi^{\mathrm{W}}(t_i) + \sigma^{\mathrm{P}}\xi^{\mathrm{P}}(t_i) \tag{F1}$$

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 $\mathrm{N} \in \{\mathrm{W}, \mathrm{P}\}$. The noise processes are normalized so that the total noise power, $\int df\; |S_{\mathrm{N}}[f]|$ is the same for any of the considered noise processes; hence the relative magnitude $(\sigma^{\mathrm{P}}/\sigma^{\mathrm{W}})^2$ determines the relative strength of the noise processes with different correlation statistics.

We restrict ourselves again to binary classification of states $|e\rangle$ and $|g\rangle$. In Fig. 10, we plot the calculated infidelities using the MF and TPP approaches against each other in logscale 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, the FGDA and 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, namely $(\sigma^{\mathrm{P}}/\sigma^{\mathrm{W}})^2 \neq 0$, and hence has a non-zero correlation timescale. We immediately note that even when the colored noise powers is only a fraction of the white noise power, the TPP-learned filters provide a non-negligible improvement over the standard MF scheme.