Editors' Suggestion

Machine Learning for Predictive Estimation of Qubit Dynamics Subject to Dephasing

Riddhi Swaroop Gupta[*] and Michael J. Biercuk

ARC Centre of Excellence for Engineered Quantum Systems, School of Physics, The University of Sydney, Sydney, New South Wales 2006, Australia

(Received 10 December 2017; published 27 June 2018)

Decoherence remains a major challenge in quantum computing hardware, and a variety of physical-layer controls provide opportunities to mitigate the impact of this phenomenon through feedback and feedforward control. In this work, we compare a variety of machine-learning algorithms derived from diverse fields for the task of state estimation (retrodiction) and forward prediction of future qubit-state evolution for a single qubit subject to classical, non-Markovian dephasing. Our approaches involve the construction of a dynamical model capturing qubit dynamics via autoregressive or Fourier-type protocols using only a historical record of projective measurements. A detailed comparison of achievable prediction horizons, model robustness, and measurement-noise-filtering capabilities for Kalman filters (KFs) and Gaussian process regression (GPR) algorithms is provided. We demonstrate superior performance from the autoregressive KF relative to Fourier-based KF approaches and focus on the role of filter optimization in achieving suitable performance. Finally, we examine several realizations of GPR using different kernels and discover that these approaches are generally not suitable for forward prediction. We highlight the linkages between predictive performance and kernel structure, and we identify ways in which forward predictions are susceptible to numerical artifacts.

DOI: 10.1103/PhysRevApplied.9.064042

I. INTRODUCTION

In predictive estimation, a dynamically evolving system is observed and any temporal correlations encoded in the observations are used to predict the future state of the system. This generic problem is well studied in diverse fields such as engineering, econometrics, meteorology, and seismology [1–5], and it is addressed in the control-theoretic literature as a form of filtering. Applying these standard classical approaches to state estimation on qubits is complicated by a variety of factors; dominant among these is the violation of the assumption of linearity inherent in most filtering applications, as qubit states are formally bilinear. The case of an idling, or freely evolving, qubit subject to dephasing is more complicated still, as an *a priori* model of system evolution suitable for implementation within standard filtering algorithms is not, in general, available.

Fortunately, there are many lessons to learn from classical control, even in the presence of such complications. For classical systems, machine-learning techniques have enabled state tracking, control, and forecasting for highly nonlinear and noisy dynamical trajectories or complex measurement protocols (e.g., Refs. [6–10]). These demonstrations move far beyond the simplified assumptions underlying many basic filtering tasks, such as linear dynamics and white (uncorrelated) noise processes. For instance, so-called particle-based Bayesian frameworks (e.g., particle, unscented or σ-point filtering) allow state estimation and tracking in the presence of nonlinearities in system dynamics or measurement protocols [11]. Further extensions approach the needs of a stochastically evolving system; recently, an ensemble of so-called unscented Kalman filters, named after the underlying mathematical procedure, demonstrated state estimation and forward predictions for chaotic, nonlinear systems in the absence of a prescribed model [10]. For nonchaotic, multicomponent stationary random signals, other algorithmic approaches have been particularly useful for tracking instantaneous frequency and phase information [12,13], enabling short-run forecasting.

In the field of quantum control, work has begun to incorporate the additional challenges faced when considering state estimation on qubits, notably quantum-state collapse under projective measurement. Under such circumstances, in which the measurement backaction strongly influences the quantum state (in contrast with the classical case), it is not straightforward to extend machine-learning predictive estimation techniques. Work to date has approached the analysis of projective measurement records on qubits as pattern recognition or image reconstruction problems, for example, in characterizing the initial or final state of a quantum system (e.g., Refs. [14–16]) or reconstructing the historical evolution of a quantum system

[*] rgup9526@uni.sydney.edu.au

based on large measurement records (e.g., Refs. [17–22]). In adaptive or sequential Bayesian learning applications, a projective measurement protocol may be designed or adaptively manipulated to efficiently yield noise-filtered information about a quantum system (e.g., Refs. [23–26]).

The demonstrations above typically assume that the object of interest either is static or stochastically evolves in a manner which is dynamically uncorrelated in time (white) as measurement protocols are repeated. This simplifying assumption falls well short of typical laboratory-based experiments, where noise processes are frequently correlated in time, and evolution may also occur rapidly relative to a measurement protocol. In such a circumstance, further complexity is introduced, as the Markov condition commonly assumed in Bayesian learning frameworks [11] is immediately violated. Even in the classical case, the problem of designing an appropriate representation of non-Markovian dynamics in Bayesian learning frameworks is an active area of research (e.g., Ref. [27]). Hence, the canonical real-time tracking and prediction problem—where a nonlinear, stochastic trajectory of a system is tracked using noisy measurements and short-run forecasts are made—is underexplored for quantum systems with projective measurements.

In this paper, we develop and explore a broad class of predictive estimation algorithms allowing us to track a qubit state undergoing *stochastic but temporally correlated* evolution using a record of projective measurements, and we forecast its future evolution. Our approaches employ machine-learning algorithms to extract temporal correlations from the measurement record and use this information to build an effective dynamical model of the system's evolution. We design a deterministic protocol to correlate Markovian processes such that a certain general class of non-Markovian dynamics can be approximately tracked without violating the assumptions of a machine-learning protocol, based on the theoretically accessible and computationally efficient frameworks of Kalman filtering (KF) and Gaussian process regression (GPR). Both frameworks provide a mechanism by which temporal correlations (equally, dynamics) are encoded into an algorithm's structure such that projection of data sets onto this structure enables meaningful learning, white-noise filtering, and effective forward prediction. We perform numerical simulations to test the effectiveness of these algorithms in maximizing the prediction horizon under various conditions, and we quantify the role of the measurement sampling rate relative to the noise dynamics in defining the prediction horizon. Simulations incorporate a variety of measurement models, including preprocessed data yielding a continuous measurement outcome and discretized outcomes commonly associated with single-shot projective qubit measurements. We find that, in most circumstances, an autoregressive Kalman framework yields the best performance, providing model-robust forward prediction horizons and effective filtering of measurement noise. Finally, we demonstrate that standard GPR-based protocols employing a variety of kernels, while effective for the problem of filtering (fitting) a measurement record, are not suitable for real-time forecasting beyond the measurement record.

In what follows, we describe in detail the physical setting for our problem in Sec. II and explain how this setting leads to a specific choice of algorithm which may be deployed for the task of tracking non-Markovian state dynamics in the absence of a dynamical model for system evolution. We provide an overview of the central GPR and KF frameworks in Sec. III, and we specify a series of algorithms under consideration in this paper tailored to different measurement processes. For preprocessed measurement records, we consider four algorithmic approaches: a least-squares filter (LSF) from Ref. [28], an autoregressive Kalman filter (AKF), a so-called Liska Kalman filter from Ref. [29] adapted for a fixed oscillator basis (LKFFB), and a suitably designed GPR learning protocol. For binary measurement outcomes, we extend the AKF to a quantized Kalman filter (QKF). In Sec. IV A, we present optimization procedures for tuning all algorithms. Numerical investigations of algorithmic performance are presented in Sec. IV, and a comparative analysis of all algorithms is provided in Sec. V.

II. PHYSICAL SETTING

Our physical setting considers a sequence of projective measurements performed on a qubit. Each projective measurement yields a 0 or 1 outcome representing the state of the qubit. The qubit is then reset, and the exact procedure is repeated. By considering a qubit state initialized in a superposition of the measurement basis (for us, Pauli $\hat{\sigma}_z$ eigenstates), we gain access to a direct probe of qubit phase evolution. If, for instance, no dephasing is present, then the probability of obtaining a binary outcome remains static in time as sequential qubit measurements are performed. If slowly drifting environmental dephasing is present, then the probability of obtaining a given binary outcome also drifts stochastically. In essence, the qubit probes dephasing noise and our procedure encodes a continuous-time non-Markovian dephasing process into time-stamped, discrete binary samples through the nonlinear projective measurement, carrying the underlying correlations in the noise. It is this series of measurements which we seek to process in our algorithmic approaches to qubit-state tracking and prediction.

Formally, an arbitrary environmental dephasing process manifests as time-dependent stochastic detuning, $\delta\omega(t)$, between the qubit frequency and the system master clock. This detuning is an experimentally measurable quantity in a Ramsey protocol, as shown schematically in Fig. 1(a). A nonzero detuning over the measurement period τ (starting from $t = 0$) induces a stochastic relative phase

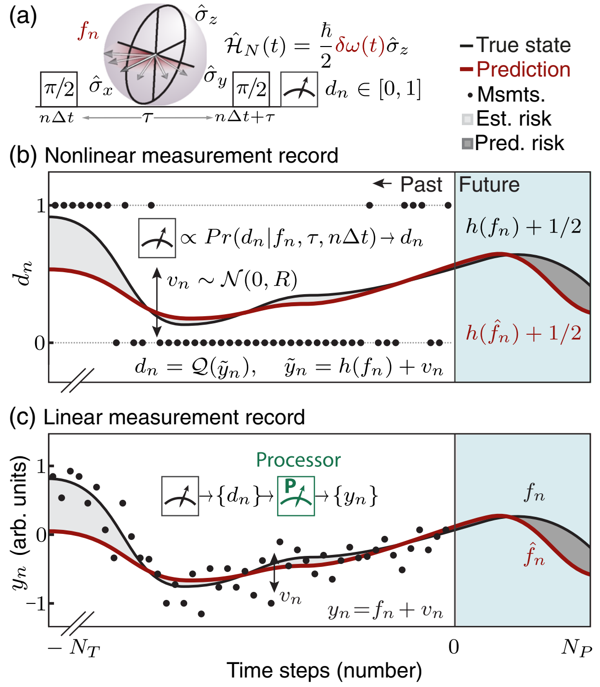
(a)
f_n
$\hat{\sigma}_z$
$\hat{\mathcal{H}}_N(t) = \frac{\hbar}{2}\delta\omega(t)\hat{\sigma}_z$
$\pi/2$
$\hat{\sigma}_x$
$\hat{\sigma}_y$
$\pi/2$
$d_n \in [0,1]$
$n\Delta t$
τ
$n\Delta t+\tau$
True state
Prediction
Msmts.
Est. risk
Pred. risk
(b) Nonlinear measurement record
Past
Future
1
0
d_n
$\propto Pr(d_n|f_n,\tau,n\Delta t) \rightarrow d_n$
$v_n \sim \mathcal{N}(0,R)$
$h(f_n)+1/2$
$h(\hat{f}_n)+1/2$
$d_n = \mathcal{Q}(\tilde{y}_n), \quad \tilde{y}_n = h(f_n)+v_n$
(c) Linear measurement record
Processor
$\rightarrow\{d_n\}\rightarrow$ P $\rightarrow\{y_n\}$
y_n (arb. units)
1
0
−1
f_n
$\hat{f}_n$
v_n
$y_n = f_n + v_n$
$-N_T$
0
N_P
Time steps (number)

FIG. 1. (a) A Ramsey experiment at $t = n\Delta t$ with fixed wait time τ and time steps n, spaced $\Delta t > \tau$ apart. A $\pi/2$ pulse rotates the qubit state to a superposition of $|d\rangle$ states, $d \in \{0, 1\}$; the qubit evolves via $\hat{\mathcal{H}}_N(t)$, accumulating relative stochastic f_n for nonzero environmental dephasing $\delta\omega(t)$. Jittering arrows depict the potential qubit-state vectors permitted for an (unknown) random f_n. The qubit state is measured as $d_n = d$ in the $\hat{\sigma}_z$ basis after a second $\pi/2$ rotation. (b) Black dots depict $\{d_n\}$ against time steps n; data collection stops at $n = 0$, separating past state estimation from future prediction (the blue region). The black solid line shows the true qubit-state likelihood $\propto h(f_n)$, and the red solid line shows the state estimate (prediction) for $n < 0$ ($n > 0$). A prediction horizon is for all $n < n^* \in [0, N_P]$, for which the dark-gray region between the red and black lines is minimized (the Bayes prediction risk) relative to predicting the mean of dephasing noise; algorithmic tuning occurs by minimizing the light-gray region (the Bayes state estimation risk). $\mathcal{Q}$ quantizes the black line into noisy qubit measurements, d_n, under the Gaussian uncertainty v_n. (c) Single-shot outcomes in (b) are preprocessed to yield noisy measurements $\{y_n\}$ (black dots); y_n is linear in f_n, and v_n represents additive white Gaussian measurement noise. Msmts. denotes measurements.

accumulation (in the rotating frame) for a qubit superposition as $|0\rangle + e^{-if(0,\tau)}|1\rangle$ between qubit basis states. The accumulated $f(0,\tau)$ at the end of a single Ramsey experiment is mapped to a probability of obtaining a particular outcome in the measurement basis via the form of the Ramsey sequence.

In a sequence of n Ramsey measurements spaced Δt apart with a fixed duration, τ, the change in the statistics of measured outcomes over this measurement record depends solely on the dephasing $\delta\omega(t)$. We assume that the measurement action over τ is much faster than the temporal dynamics of the dephasing process, and that $\Delta t \gtrsim \tau$. The resulting measurement record is a set of binary outcomes, $\{d_n\}$, determined probabilistically from n true stochastic qubit phases, $f \coloneqq \{f_n\}$. Here the accumulated phase in each Ramsey experiment $f(n\Delta t, n\Delta t+\tau) \equiv \int_{n\Delta t}^{n\Delta t+\tau} \delta\omega(t')dt'$, and we use the shorthand $f(n\Delta t, n\Delta t + \tau) \equiv f_n$. We define the statistical likelihood for observing a single shot, d_n, using Born's rule [30]:

$$\Pr(d_n = d|f_n, \tau, n\Delta t) = \begin{cases} \cos^2\left(\frac{f_n}{2}\right) & \text{for } d = 1 \\ \sin^2\left(\frac{f_n}{2}\right) & \text{for } d = 0 \end{cases}. \quad (1)$$

The notation $\Pr(d_n|f_n, \tau, n\Delta t)$ refers to the conditional probability of obtaining measurement outcome d_n given a true stochastic phase, f_n, accumulated over τ, beginning at time $t = n\Delta t$. As an example, in the noiseless case, $\Pr(d_n = 1|f_n, \tau, n\Delta t) = 1$, $\forall\, n$, such that a qubit exhibits no additional phase accumulation due to environmental dephasing. In general, after a measurement at n, the qubit state is reset, but the dephasing noise correlations manifest again via Born's rule for another random value of the bias at the time step $n + 1$. A detailed discussion of Eq. (1) can be found in Appendix A.

The action of measurement, expressed as $h(f_n)$, is given by $\Pr(d_n = d|f_n, \tau, n\Delta t) \equiv \frac{1}{2} - (-1)^d h(f_n)$ and is depicted in Fig. 1(b) as a probability of seeing the qubit in the $d = 1$ state. We begin by describing here a "raw" nonlinear measurement record, $\{d_n\}$ where each d_n (indicated by a black dot) corresponds to a binary outcome derived from a single projective measurement on a qubit. The sequence $\{d_n\}$ can be treated as a sequence of biased coin flips, where the underlying bias of the coin is a non-Markovian, discrete-time process and the value of the bias is given by Eq. (1) at each n. The nonlinearity of the measurement, $h(f_n)$, is defined with respect to f_n, where Eq. (1) is interpreted as a nonlinear measurement action for Bayesian learning frameworks.

This data series is contrasted with a linear measurement record, $\{y_n\}$, depicted in Fig. 1(c). Each value y_n is derived from the sum of a true qubit phase, f_n, and Gaussian white measurement noise, v_n. The sequence $\{y_n\}$ is generated by preprocessing raw binary measurements, $\{d_n\}$, via a range of experimental techniques subject to a separation of timescales about τ, such that τ is much faster than the drift of $\delta\omega(t)$. In the most common case, one performs M runs of the experiment over which $\delta\omega(t)$ is approximately constant, giving an estimate of f_n at $t = n\Delta t$ using averaging, a Bayesian scheme, or Fourier analysis. A more complex linearization protocol involves the use of low-pass or decimation filtering on a sequence $\{d_n\}$ to yield $\hat{\Pr}(d_n|f_n, \tau, n\Delta t)$, from which the accumulated phase corrupted by measurement noise, $\{y_n\}$, can be obtained from Eq. (1). Since any low-pass or decimation filter has an averaging effect on a signal, decimation filtering a sequence $\{d_n\}$ provides an alternative, software-based approach to physically averaging single-shot qubit measurements. Hence, the linear measurement record in Fig. 1(c) arises either from software preprocessing (filtering) data from a single qubit or from experimental averaging over an ensemble of qubits.

We impose properties on environmental dephasing such that our theoretical designs can enable meaningful predictions. We assume that dephasing is non-Markovian, covariance stationary, and mean-square ergodic. That is, a single realization of the process f is drawn from a power spectral density of arbitrary, but non-Markovian, form. We further assume that f is a Gaussian process and that the separation of timescales between measurement protocols and dephasing dynamics articulated above are met.

Given these conditions, our task is to build a dynamical model to approximately track f over past measurements ($n < 0$) and enable qubit-state predictions in future times ($n > 0$). This prediction is represented by the red lines in Figs. 1(b) and 1(c), and it differs from the truth by the so-called estimation (prediction) risk for past (future) times as indicated by the shading. We represent our estimate of f for all times using a hat in both the linear and nonlinear measurement models. The major challenge we face in developing this estimate, $\hat{f}$ [equivalently, $\hat{\Pr}(d_n|f_n, \tau, n\Delta t)$], is that, for a qubit evolving under stochastic dephasing [the true state given by the black solid lines in Figs. 1(b) and 1(c)], we have no *a priori* dynamical model for the underlying evolution of f. In the next section, we define the theoretical structures of KF and GPR algorithms which allow us to build that dynamical model directly from the historical measurement record.

III. OVERVIEW OF PREDICTIVE METHODOLOGIES

Our objective is to implement an algorithm permitting learning of the underlying qubit dynamics in such a way as to maximize the forward prediction horizon for a given qubit data record. We first quantify the quality of our state estimation procedure. The fidelity of any underlying algorithm during state estimation and prediction, relative to the true state, is expressed by the mathematical quantity known as the Bayes risk, where zero risk corresponds to a perfect estimation. At each time step n, the Bayes risk is a mean-square distance between the truth, f, and the prediction, $\hat{f}$, calculated over an ensemble of M different realizations of true f and noisy data sets $\mathcal{D}$:

$$L_{\rm BR}(n|I) \equiv \langle (f_n - \hat{f}_n)^2 \rangle_{f,\mathcal{D}}. \qquad (2)$$

The notation $L_{\rm BR}(n|I)$ expresses that the Bayes risk value at n is conditional on I, a placeholder for free parameters in the design of the predictor, $\hat{f}_n$. State estimation risk is the Bayes risk incurred during $n \in [-N_T, 0]$; prediction risk is the Bayes risk incurred during $n \in [0, N_P]$. The state estimation and prediction risk regions for one realization of dephasing noise are shaded in Figs. 1–3. We therefore define the forward prediction horizon as the number of time steps for $n^* \in [0, N_P]$ during which a predictive algorithm incurs a lower Bayes prediction risk than naively predicting $\hat{f}_n \equiv \mu_f = 0 \; \forall \; n$, the mean qubit behavior under zero-mean dephasing noise.

With this concept in mind, we introduce two general approaches for algorithmic learning relevant to the strictures of the problem we have introduced. Our general approach is shared between all algorithms employed and is represented schematically for the KF and GPR in Fig. 2. Stochastic qubit evolution is depicted for one realization of f (the black solid line) given noisy linear measurements (the black dots) corrupted by the Gaussian white measurement noise v_n. Our overall task is to produce an estimate, given by the red line, which minimizes risk for the prediction period. Ideally, both estimation risk and prediction risk are minimized simultaneously for well-performing implementations.

Examining the insets in both panels of Fig. 2, both frameworks start with a prior Gaussian distribution over qubit states (purple) that is constrained by the measurement record to yield a posterior Gaussian distribution of the qubit state (red). The prior captures assumptions about the qubit state before any data are seen and the posterior expresses our best knowledge of the qubit state under a Bayesian framework. The posterior distribution in both KF and GPR is used to generate qubit-state estimates ($n < 0$) and predictions ($n > 0$) (the red solid line). However, the computational process by which this posterior is inferred differs significantly between the two methods; we provide an overview of the central features of these algorithms below.

The key feature of a Kalman filter is the recursive learning procedure shown in the inset to Fig. 2(a). Our knowledge of the qubit state is summarized by the prior and posterior Gaussian probability distributions, and they are created and collapsed recursively *at each time step*. The mean of these distributions is the true Kalman state, x_n, and the covariance of these distributions, P_n, captures the uncertainty in our knowledge of x_n; together, both define the Gaussian distribution. The Kalman filter produces an *estimate* of the state, $\hat{x}_n$, at each step through this recursive procedure, taking into account two factors. First, the Kalman gain, γ_n, updates our knowledge of (x_n, P_n) within each time step n and serves as a weighting factor for the difference between incoming data, and our best estimate for an observation based on $\hat{x}_n$, suitably transformed via the measurement action, $h(\hat{x}_n)$. Next, the dynamical model Φ_n propagates the state and covariance, (x_n, P_n), to the next time step, such that the posterior moments at n define the prior at $n + 1$. This process occurs for each time step, and an estimate of a true x_n state is built up recursively based on

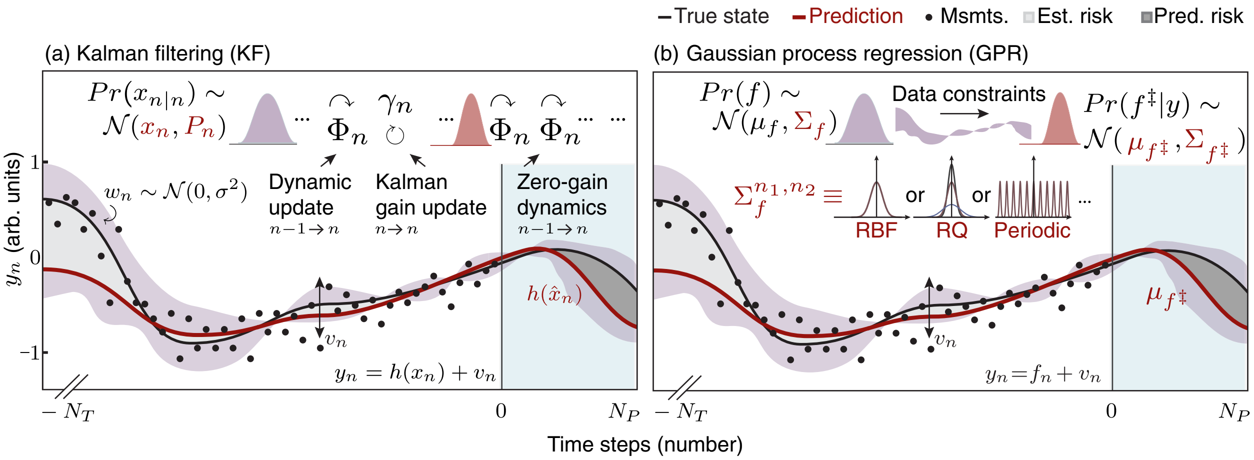
True state
Prediction
Msmts.
Est. risk
Pred. risk
(a) Kalman filtering (KF)
$Pr(x_{n|n}) \sim \mathcal{N}(x_n, P_n)$
Φ_n
γ_n
Φ_n Φ_n
$w_n \sim \mathcal{N}(0,\sigma^2)$
Dynamic update $n-1 \rightarrow n$
Kalman gain update $n \rightarrow n$
Zero-gain dynamics $n-1 \rightarrow n$
$h(\hat{x}_n)$
v_n
$y_n = h(x_n) + v_n$
y_n (arb. units)
1
0
−1
$-N_T$
0
N_P
(b) Gaussian process regression (GPR)
$Pr(f) \sim \mathcal{N}(\mu_f, \Sigma_f)$
Data constraints
$Pr(f^\ddagger|y) \sim \mathcal{N}(\mu_{f^\ddagger}, \Sigma_{f^\ddagger})$
$\Sigma_f^{n_1,n_2} \equiv$
RBF
or
RQ
or
Periodic
$\mu_{f^\ddagger}$
v_n
$y_n = f_n + v_n$
$-N_T$
0
N_P
Time steps (number)

FIG. 2. Comparison of the algorithmic structure between KF and GPR by superposing the lower panels of Fig. 1 with KF and GPR predictive frameworks. (a) KF: Purple distribution represents a prior, with mean x_n and covariance P_n, propagated in time steps n, using Kalman dynamics Φ_n, and updated within each n by the Kalman gain γ_n to yield a posterior distribution (red) at n. The posterior at n is the prior at $n+1$. The mean of a posterior distribution at each n is used to derive predictions given by the red line using $h(x_n)$. In the blue region, the red posterior predictive distribution is propagated using Φ_n, but $\gamma_n \equiv 0$. Gaussian white Kalman "process" noise, w_n, is colored by Φ_n to yield the dynamics for x_n. (b) Purple prior distribution defined over sequences f, with mean μ_f, and variance Σ_f is constrained by the entire measurement record. The resulting posterior predictive distribution (red) is evaluated at test points in time, $n^\ddagger \in [-N_T, N_P]$; state estimates (predictions) are for the mean, $\mu_{f^\ddagger}$ at $n^\ddagger < 0$ ($n^\ddagger > 0$). A choice of kernel defines each element in Σ_f, $\Sigma_{f^\dagger}$. In both (a) and (b), the purple shadow represents posterior state variance (the diagonal P_n or $\Sigma_{f^\ddagger}$ elements) constrained by data and filtered measurement noise v_n.

all of our existing knowledge—namely, a linear combination of all past measurements—and all previously generated state estimates. Beyond $n = 0$, we perform predictions in the absence of further measurement data by simply propagating the dynamic model with the Kalman gain set to zero. Full details of the KF algorithm appear in Sec. III A.

In our application, we define the Kalman state, x_n, the dynamical model Φ_n, and a measurement action $h(x_n)$ such that the Kalman filtering framework can track a non-Markovian qubit-state trajectory due to an arbitrary realization of f. In standard KF implementations, the discrete-time sequence $\{x_n\}$ defines a "hidden" signal that cannot be observed, and the dynamic model Φ_n is known. We deviate from this standard construction such that our true Kalman state and its uncertainty, (x_n, P_n), do not have a direct physical interpretation. The Kalman x_n has no deterministic component and corresponds to arbitrary power spectral densities describing f. Hence, the role of the Kalman x_n is to represent an abstract correlated process that, upon measurement, yields physically relevant quantities governing qubit dynamics. Moreover, a key challenge described in detail below is to construct an effective Φ_n from the measurement record.

In contrast to the recursive approach taken in the KF, a GPR learning protocol illustrated schematically in Fig. 2(b) selects a *random process* to best describe the overall dynamical behavior of the qubit state under one realization of f. The key point is that sampling the prior or posterior distribution in GPR yields random realizations of discrete time *sequences*, rather than individual random variables, and GPR considers the entire measurement record at once. In a sense, it corresponds to a form of fitting over the entire data set. The output of a GPR protocol is a predictive distribution which we can evaluate at an arbitrarily chosen sequence of test points, where the test points can exist for $n < 0$ ($n > 0$), such that we extract state estimates (forward predictions) from the predictive distribution. Owing to the nature of this procedure, we wish to distinguish the set of test points (in units of time steps) by using a double dagger, namely, that we are evaluating the predictive posterior distribution of a GPR protocol at the desired time labels. In this notation, $\{n^\ddagger\}$, $n^\ddagger \in [-N_T, N_P]$ are the test points and $N^\ddagger$ is the total length of an array of test points, where a state estimation occurs if $n^\ddagger \leq 0$ and predictions occur if $n^\ddagger > 0$.

The process of building the posterior distribution is implemented using a kernel, or basis, from which to construct the effective fit. In standard GPR implementations, the correlation between any two observations depends only on the separation distance of the index of these observations, and correlations are captured in the covariance matrix, Σ_f. Each element $\Sigma_f^{n_1,n_2}$ describes this correlation for observations at arbitrary time steps indexed by n_1 and n_2: this quantity is given in a form set by the selected kernel.

In our application, the non-Markovian dynamics of f are not specified explicitly but are encoded in a general way through the choice of kernel, prescribing how $\Sigma_f^{n_1,n_2}$ should be calculated. The Fourier transform of the kernel represents a power spectral density in Fourier space. A general design of

$\Sigma_f^{n_1,n_2}$ allows one to probe arbitrary stochastic dynamics and, equivalently, explore arbitrary regions in the Fourier domain. For example, Gaussian kernels (RBFs) and mixtures of Gaussian kernels (RQs) capture the continuity assumption that correlations die out as the separation in time increases. We choose to employ an infinite basis of oscillators implemented by the so-called periodic kernel to enable us to represent arbitrary power spectral densities for f. A prediction occurs simply by extending the GPR fit by choosing test points $n^\ddagger > 0$.

In the following subsections, we provide details of the specific classes of the learning algorithm employed here, with an eye towards evaluating their predictive performance on qubit-measurement records. We introduce a series of KF algorithms capable of handling both linear and nonlinear measurement records, and we restrict our analysis of GPR to linear measurement records.

A. Kalman framework

In order for a Kalman filter to track a stochastically evolving qubit state in our application, the hidden true Kalman state at time step n, x_n, must mimic stochastic dynamics of a qubit under environmental dephasing. We propagate the hidden state x_n according to a dynamical model Φ_n corrupted by Gaussian white process noise, w_n:

$$x_n = \Phi_n x_{n-1} + \Gamma_n w_n, \tag{3}$$

$$w_n \sim \mathcal{N}(0, \sigma^2) \quad \forall\, n. \tag{4}$$

Process noise has no physical meaning in our application —w_n is shaped by Γ_n and deterministically colored by the dynamical model Φ_n to yield a non-Markovian x_n representing qubit dynamics under generalized environmental dephasing. In addition to coloring via the dynamical model, one can shape input white noise by designing Γ_n.

We measure x_n using an ideal measurement protocol, $h(x_n)$, and incur additional Gaussian white measurement noise v_n with scalar covariance strength R, yielding scalar noisy observations y_n:

$$y_n = z_n + v_n, \tag{5}$$

$$z_n \equiv h(x_n), \tag{6}$$

$$v_n \sim \mathcal{N}(0, R) \quad \forall\, n. \tag{7}$$

The measurement procedure, $h(x_n)$, can be linear or nonlinear, allowing us to explore both regimes in our physical application.

With appropriate definitions, the Kalman equations below specify all Kalman algorithms in this paper. At each time step n, we denote estimates of the moments of the prior and posterior distributions (equivalently, estimates of the true Kalman state) with $(\hat{x}_n(-), \hat{P}_n(-))$ and $(\hat{x}_n(+), \hat{P}_n(+))$ respectively. The Kalman update equations take a generic form (e.g., see Ref. [31]):

$$\hat{x}_n(-) = \Phi_{n-1}\hat{x}_{n-1}(+), \tag{8}$$

$$Q_{n-1} = \sigma^2 \Gamma_{n-1}\Gamma_{n-1}^T, \tag{9}$$

$$\hat{P}_n(-) = \Phi_{n-1}\hat{P}_{n-1}(+)\Phi_{n-1}^T + Q_{n-1}, \tag{10}$$

$$\gamma_n = \hat{P}_n(-)H_n^T[H_n\hat{P}_n(-)H_n^T + R_n]^{-1}, \tag{11}$$

$$\hat{y}_n(-) = h(\hat{x}_n(-)), \tag{12}$$

$$\hat{x}_n(+) = \hat{x}_n(-) + \gamma_n(y_n - \hat{y}_n(-)), \tag{13}$$

$$\hat{P}_n(+) = [1 - \gamma_n H_n]\hat{P}_n(-). \tag{14}$$

To reiterate, Eqs. (8) and (10) bring the best state of knowledge from the previous time step into the current time step n, as a prior distribution. Dynamical evolution is modified by features of the process noise, as encoded in Eq. (9) and propagated in Eq. (10). The propagation of the moments of the prior distribution, as outlined thus far, does not depend on the incoming measurement, y_n, but is determined entirely by the dynamical model—in our case, $\Phi \equiv \Phi_n$, $\forall\, n$.

The Kalman gain in Eq. (11) depends on the uncertainty in the true state, $\hat{P}_n(-)$ and is modified by features of the measurement model, H_n, and measurement noise, $R_n \equiv R$, $\forall\, n$. It serves as an effective weighting function for each incoming observation. Before seeing any new measurement data, the filter predicts an observation $\hat{y}_n(-)$ corresponding to the best available knowledge at n in Eq. (12). This value is compared to the actual noisy measurement y_n received at n, and the difference is used to update our knowledge of the true state via Eq. (13). If measurement data are noisy and unreliable (a high R value), then γ has a small value, and the algorithm propagates Kalman state estimates according to the dynamical model and effectively ignores the data. In particular, only the second terms in both Eq. (13) and Eq. (14) represent the Bayesian update of the moments of a prior distribution [the $(-)$ terms] to the posterior distribution [the $(+)$ terms] at n. If $\gamma_n \equiv 0$, then the prior and posterior moments at any time step are exactly identical by Eqs. (13) and (14), and only dynamical evolution occurs using Eqs. (8)–(10). This is the condition we employ when we seek to make forward predictions beyond a single time step, and hence we set $\gamma \equiv 0$ during future prediction.

Since we do not have a known dynamical model Φ for describing stochastic qubit dynamics under f, we need to make design choices for $\{x, \Phi, h(x), \Gamma\}$ such that f can be approximately tracked. These design choices completely specify the algorithms introduced below and represent key findings with respect to our work in this paper. For a linear measurement record, $h(x) \mapsto Hx$, and we compare the predictive performance for Φ modeling stochastic dynamics either via so-called autoregressive processes in the AKF or via projection onto a collection of oscillators in the LKFFB. In addition, we use the dynamics of an AKF to define a QKF with a nonlinear, quantized measurement model such that the filter can act directly on binary qubit outcomes. We provide the relevant details in the subsections below.

1. AKF

Recursive autoregressive methods are well studied in classical control applications (cf. Ref. [32]) presenting opportunities to leverage existing engineering knowledge in developing quantum control strategies. In our application, we use an autoregressive Kalman filter to probe arbitrary, covariance-stationary qubit dynamics such that the dynamic model is constructed as a weighted sum of q past values driven by white noise, i.e., an autoregressive process of order q, AR(q). Using Wold's decomposition, it can be shown that any zero-mean covariance-stationary process representing qubit dynamics has a representation in the mean-square limit by an autoregressive process of finite order, as in Appendix B.

The study of AR(q) processes falls under a general class of techniques based on autoregressive moving average (ARMA) models in adaptive control engineering and econometrics (see, e.g., Refs. [33,34], respectively). For high-q models in a typical time-series analysis, it is possible to decompose an AR(q) into an ARMA model with a small number of parameters [35,36]. However, we retain a high-q model to probe arbitrary power spectral densities. Furthermore, the literature suggests that employing a high-q model is relatively easier than a full ARMA estimation problem and enables lower prediction error [35,37].

To construct the Kalman dynamical operator Φ for the AKF, we introduce a set of q coefficients $\{\phi_{q' \leq q}\}$, $q' = 1, \ldots, q$ to specify the dynamical model:

$$f_n = \phi_1 f_{n-1} + \phi_2 f_{n-2} + \cdots + \phi_q f_{n-q} + w_n. \qquad (15)$$

We thus see that the dynamical model is constructed as a weighted sum of time-retarded samples of f, with weighting factors given by the autoregressive coefficients up to order (and hence time lag) q. For small values of $q < 3$, it is possible to extract simple conditions on the coefficients, $\{\phi_{q' \leq q}\}$, that guarantee properties of f: for example, that f is covariance stationary and mean-square ergodic. In our application, we freely employ arbitrary-q models via machine learning in order to improve our approximation of an arbitrary f. Any AR(q) process can be recast (nonuniquely) into state space form [4], and we define the AKF by the following substitutions into Kalman equations:

$$x_n \equiv [f_n, \ldots, f_{n-q+1}]^T, \qquad (16)$$

$$\Gamma_n w_n \equiv [w_n, 0, \ldots, 0]^T, \qquad (17)$$

$$\Phi_{\mathrm{AKF}} \equiv \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{q-1} & \phi_q \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad \forall n, \qquad (18)$$

$$H \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \quad \forall n. \qquad (19)$$

The matrix Φ_{AKF} is the dynamical model used to recursively propagate the unknown state during state estimation in the AKF, as represented schematically in the upper half of Fig. 3. In general, the $\{\phi_{q' \leq q}\}$ employed in Φ_{AKF} must be learned through an optimization procedure using the measurement record, where the set of parameters to be optimized is $\{\phi_1, \ldots, \phi_q, \sigma^2, R\}$. This procedure yields the

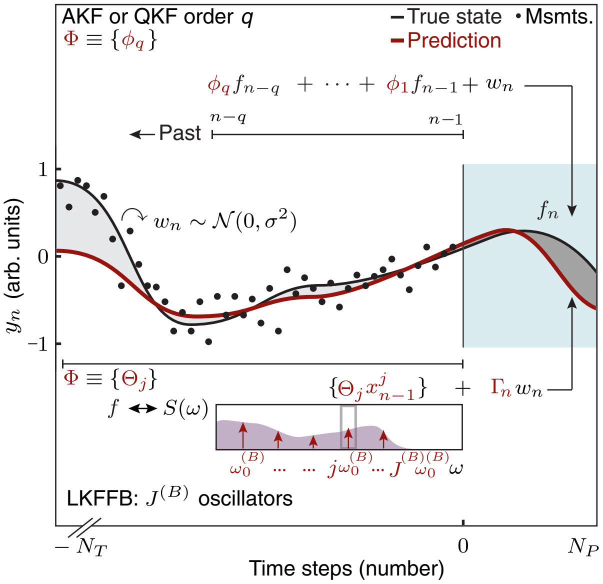
AKF or QKF order q
$\Phi \equiv \{\phi_q\}$
True state
Msmts.
Prediction
$\phi_q f_{n-q} + \cdots + \phi_1 f_{n-1} + w_n$
Past
$n-q$
$n-1$
$w_n \sim \mathcal{N}(0, \sigma^2)$
f_n
1
0
−1
y_n (arb. units)
$\Phi \equiv \{\Theta_j\}$
$\{\Theta_j x_{n-1}^j\}$
+
$\Gamma_n w_n$
$f \leftrightarrow S(\omega)$
$\omega_0^{(B)}$ … … $j\omega_0^{(B)}$ … $J^{(B)}\omega_0^{(B)}$ ω
LKFFB: $J^{(B)}$ oscillators
$-N_T$
0
N_P
Time steps (number)

FIG. 3. Approaches to construction of the KF dynamical model. Figure 2(a) is superimposed with Kalman dynamical models, $\Phi \equiv \Phi_n, \forall\, n$. (a) AKF or QKF. A set of autoregressive coefficients, $\{\phi_{q' \leq q}\}$, define Φ to yield f_n as a weight sum of q past measurements. (b) LKFFB. Red arrows with heights $\|x_n^j\|$ depict a set of basis oscillators for $j = 1, \ldots, J^{(B)}$, probe the true purple spectrum of f_n, and yield time-domain dynamics of f_n as a stacked system of resonators, Θ_j. Black L-shaped arrows depict a single instance of f_n at $n = 0$ based on historical $\{f_{n-1}, f_{n-2}, \cdots\}$ values.

optimal configuration of the autoregressive Kalman filter, but at the computational cost of a $(q+2)$-dimensional Bayesian learning problem for an arbitrarily large q.

The LSF in Ref. [28] considers a weighted sum of past measurements to predict the ith-step-ahead measurement outcome, $i \in [0, N_P]$. A gradient descent algorithm learns the weights $\{\phi_{q' \leq q}\}$ for the previous q past measurements, and a constant offset value for nonzero mean processes is used to calculate the ith-step-ahead prediction. The set of N_P LSF models, taken collectively, define the set of predicted qubit states under a LSF acting on a measurement record. For $i = 1$, equivalent to the single-step update employed in the Kalman filter, we assert that the learned $\{\phi_{q' \leq q}\}$ in the LSF effectively implements an AR(q) process (which we validate numerically in Sec. IV). Under this condition, and for zero-mean w_n, the LSF in Ref. [28], by definition, searches for coefficients for the weighted linear sum of past q measurements, as described in Eq. (15).

We use the parameters $\{\phi_{q' \leq q}\}$ learned in the LSF to define Φ in Eq. (18), therefore reducing the computational complexity of the remaining optimization from $[(q+2) \rightarrow 2]$ dimensional for an AKF of order q. Since Kalman noise parameters (σ^2, R) are subsequently autotuned using a Bayes-risk optimization procedure (see Sec. IV A), we optimize over the potential remaining model errors and measurement noise.

In general, LSF performance improves as q increases, and a full characterization of the model-selection decisions for the LSF is given in Ref. [28]. Defining an absolute value for the optimal q is somewhat arbitrary, as it is defined relative to the extent to which a true f is oversampled in the measurement routine and the finite size of the data. For all analyses presented here, we fix the ratio at $q\Delta t = 0.1$ (arb. units) and $q/N_T = 0.05$ (arb. units), where the experimental sampling rate is $1/\Delta t$, and N_T and $\{\phi_{q' \leq q}\}$ are identical in the AKF and the LSF. In practice, these ratios ensure numerical convergence of the LSF during training.

2. LKFFB

In LKFFB, we effectively perform a Fourier decomposition of the underlying f in order to build the dynamic model, Φ, for the Kalman filter. Here, we project our measurement record on $J^{(B)}$ oscillators with a fixed frequency $\omega_j \equiv j\omega_0^{(B)}$, with j an integer, $j = 1, \ldots, J^{(B)}$. The temporal resolution of the state-tracking procedure is set by the maximum frequency in the selected basis and the properties of the spacing between adjacent basis frequencies. The superscript $^{(B)}$ indicates Fourier-domain information about an algorithmic basis, as opposed to information about the true (unknown) dephasing process. The LKFFB allows instantaneous amplitude and phase tracking for each basis oscillator, directly enabling forward prediction from the learned dynamics. The structure of this Kalman filter, referred to as the LKF, was developed in Ref. [29]; adding a fixed basis in this application yields the LKFFB.

For our application, the true hidden Kalman state, x, is encoded as a collection of substates, x^j, for the jth oscillator. For clarity, we remind the reader that the superscript is used as an index rather than a power. Each substate is labeled by a real and imaginary component which we represent in vector notation:

$$x_n \equiv [x_n^1, \ldots, x_n^j, \ldots, x_n^{J^{(B)}}], \tag{20}$$

$$A_n^j \equiv \mathrm{Re}(x_n^j), \tag{21}$$

$$B_n^j \equiv \mathrm{Im}(x_n^j), \tag{22}$$

$$x_n^j \equiv \begin{bmatrix} A_n^j \\ B_n^j \end{bmatrix}. \tag{23}$$

The algorithm tracks the real and imaginary parts of the Kalman substate simultaneously in order calculate the instantaneous amplitudes $(\|x_n^j\|)$ and phases (θ_n^j) for each Fourier component:

$$\|x_n^j\| \equiv \sqrt{(A_n^j)^2 + (B_n^j)^2}, \tag{24}$$

$$\theta_n^j \equiv \tan \frac{B_n^j}{A_n^j}. \tag{25}$$

The dynamical model for the LKFFB is now constructed as a stacked collection of these independent oscillators. The substate dynamics match the formalism of a Markovian stochastic process defined on a circle for each basis frequency, ω_j, as in Ref. [38]. We stack $\Theta(j\omega_0^{(B)}\Delta t)$ for all ω_j values along the diagonal to obtain the full dynamical matrix for Φ_n:

$$\Phi_n \equiv \begin{bmatrix} \Theta(\omega_0^{(B)}\Delta t) \cdots 0 \\ \cdots \Theta(j\omega_0^{(B)}\Delta t) \cdots \\ 0 \cdots \Theta(J^{(B)}\omega_0^{(B)}\Delta t) \end{bmatrix}, \tag{26}$$

$$\Theta(j\omega_0^{(B)}\Delta t) \equiv \begin{bmatrix} \cos(j\omega_0^{(B)}\Delta t) & -\sin(j\omega_0^{(B)}\Delta t) \\ \sin(j\omega_0^{(B)}\Delta t) & \cos(j\omega_0^{(B)}\Delta t) \end{bmatrix}. \tag{27}$$

We obtain a single estimate of the true hidden state by defining the measurement model, H, by concatenating $J^{(B)}$ copies of the row vector [10]:

$$H \equiv [10 \cdots 10 \cdots 10]. \tag{28}$$

Here, the unity values of H pick out and sum the Kalman estimate for the real components of f while ignoring the imaginary components; namely, we sum A_n^j for all $J^{(B)}$ basis oscillators.

In Ref. [29], a state-dependent process-noise-shaping matrix is introduced to enable potentially nonstationary instantaneous amplitude tracking in the LKFFB for each individual oscillator:

$$\Gamma_{n-1} \equiv \Phi_{n-1} \frac{x_{n-1}}{\|x_{n-1}\|}. \tag{29}$$

For the scope of this paper, we retain the form of Γ_n in our application even if true qubit dynamics are covariance stationary. As such, Γ_n depends on the state estimates x. For this choice of Γ_n, we deviate from classical Kalman filters because recursive equations for P cannot be propagated in the absence of measurement data. Consequently, Kalman gains cannot be precomputed prior to experimental data collection. Details of gain precomputation in classical Kalman filtering can be found in standard textbooks (see, e.g., Ref. [31]).

There are two ways to conduct forward prediction for a LKFFB and both are numerically equivalent for an appropriate choice of basis: (i) we set the Kalman gain to zero and recursively propagate using Φ, and (ii) we define a harmonic sum using the basis frequencies and the learned $\{\|x_n^j\|, \theta_n^j\}$. This harmonic sum can be evaluated for all future times to yield forward predictions in a single calculation. The choice of basis for a LKFFB and its implications for optimal predictive performance are discussed in Appendix C 2.

3. QKF

In a QKF, we implement a Kalman filter that acts directly on discretized measurement outcomes, $d \in \{0, 1\}$. To reiterate the discussion of Fig. 1(a), this means that the measurement action in a QKF must be nonlinear and take as input quantized measurement data. In our application, we set the dynamical model to be identical to that employed in the AKF, allowing isolation of the effect of the nonlinear, quantized measurement action.

With unified notation across the AKF and QKF, we define a nonlinear measurement model $h(x)$ and its Jacobian, H, as

$$z_n \equiv h(x_n[0]) \equiv \frac{1}{2}\cos(f_n), \tag{30}$$

$$\Rightarrow H_n \equiv \frac{dh(f_n)}{df_n} = -\frac{1}{2}\sin(f_n). \tag{31}$$

During filtering, $z_n = h(x_n[0])$ is used to compute measurement residuals when updating the true Kalman state, x_n, whereas the state variance estimate, P_n, is propagated using the Jacobian, H_n. Furthermore, the Jacobian is used to compute the Kalman gain. Hence, the filter can quickly destabilize if the linearization of $h(\cdot)$ by H_n does not hold during dynamical propagation, resulting in a rapid buildup of errors.

In this construction, the entity z_n is associated with an abstract "signal": a sequence formed by repeated applications of the likelihood function for the single-qubit measurements in Eq. (1). The true stochastic qubit phase, f_n, is our Kalman hidden state, x_n. Subsequently, we extract an estimate of the true bias, z_n, as an unnatural association of the Kalman measurement model with Born's rule. The sequence $\{z_n\}$ is not observable but can be inferred only over a large number of experimental runs.

To complete the measurement action, we implement a biased coin flip within the QKF filter given $\tilde{y}_n$. While the qubit provides measurement outcomes which are naturally quantized, we require a theoretical model, $\mathcal{Q}$, to generate quantized measurement outcomes with statistics that are consistent with Born's rule in order to propagate the dynamic Kalman filtering equations appropriately. In order to build this machinery, we modify the procedure in Ref. [39] to quantize z_n using biased coin flips. In our notation, we represent a black-box quantizer, $\mathcal{Q}$, that gives only a 0 or a 1 outcome based on $\tilde{y}_n$:

$$d_n = \mathcal{Q}(\tilde{y}_n) \tag{32}$$

$$= \mathcal{Q}(h(f_n) + v_n). \tag{33}$$

The use of the notation $\tilde{y}_n$ is meant to indicate a correspondence with y_n introduced earlier, while the physical meaning differs due to the discretized nature of the QKF. Therefore, the stochastic changes in $\{\tilde{y}_n\}$ are represented in the bias of a coin flip, subject to proper normalization constraints, which maintains $|\tilde{y}_n| \leq 0.5$:

$$\Pr(d_n|\tilde{y}_n, f_n, \tau) \equiv \mathcal{B}(n_{\mathcal{B}} = 1; p_{\mathcal{B}} = \tilde{y}_n + 0.5). \tag{34}$$

QKF uses Eq. (34) to define a biased coin flip during filtering, where $n_{\mathcal{B}}$ represents a single coin flip, and $p_{\mathcal{B}}$ represents the stochastically drifting bias on the coin. Kalman filtering with the coin-flip quantization defined by Eq. (34) presents a departure from the classical amplitude quantization procedures in Refs. [39,40].

From a computational perspective, we modify the process-noise-feature definition from an AKF to a QKF. We set $Q \equiv \sigma^2\Gamma\Gamma^T \to \sigma^2\mathcal{I}\ \forall\, n$, and $\mathcal{I}$ is a $q \times q$ identity matrix from the AKF to the QKF. The rationale for this modification is that it smears out the effect of white process noise in a way that stabilizes inversions in the gain calculation in the Kalman filter, but it does not correlate any two Kalman states in time (a diagonal matrix). In practice, this modification yields only mild improvements over the original AKF process-noise-feature matrix.

The definitions of $\{\mathcal{Q}, h(x_n), H_n, Q\}$ in this subsection and dynamics $\{x_n, \Phi\}$ from the AKF now completely specify the QKF algorithm for application to a discrete, single-shot measurement record, as depicted in Fig. 1(a).

B. GPR

In GPR, correlations in the measurement record can be learned if one projects data on a distribution of Gaussian processes, $\Pr(f)$, with an appropriate encoding of their covariance relations via a kernel, $\Sigma_f^{n_1,n_2}$. We return to the linear measurement record and the definition of scalar noisy observations y_n corrupted by Gaussian measurement noise, v_n, as considered previously for the AKF, LSF, and LKFFB. Under linear operations, the distribution of measured outcomes, y_n, is also a Gaussian. The mean and variance of $\Pr(y)$ depends on the mean μ_f and variance Σ_f of the prior $\Pr(f)$, and the mean $\mu_v \equiv 0$ and variance R of the measurement noise:

$$f \sim \Pr_f(\mu_f, \Sigma_f), \tag{35}$$

$$y \sim \Pr_y(\mu_f, \Sigma_f + R). \tag{36}$$

For covariance stationary f, correlation relationships depend solely on the time lag $\nu \equiv \Delta t|n_1 - n_2|$ between any two time points $n_1, n_2 \in [-N_T, N_P]$. An element of the covariance matrix, $\Sigma_f^{n_1,n_2}$, corresponds to one value of lag, ν, and the correlation for any given ν is specified by the covariance function, $R(\nu)$:

$$\Sigma_f^{n_1,n_2} \equiv R(\nu). \tag{37}$$

Any unknown parameters in the encoding of correlation relations via $R(\nu)$ are learned by solving the optimization problem outlined in Sec. IV A. The optimized GPR model is then applied to data sets corresponding to different realizations of f. Let the indices $n \in N_T \equiv [-N_T, 0]$ denote training points, and let a length $N^\ddagger$ vector contain the arbitrary testing points $n^\ddagger \in [-N_T, N_P]$. These testing points in machine-learning language encompass both state estimation and prediction points in our notation. We now define the joint distribution $\Pr(y, f^\ddagger)$, where $f^\ddagger$ represents the true process evaluated by GPR at the desired test points:

$$\begin{bmatrix} f^\ddagger \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{f^\ddagger} \\ \mu_y \end{bmatrix}, \begin{bmatrix} K(N^\ddagger, N^\ddagger) & K(N_T, N^\ddagger) \\ K(N^\ddagger, N_T) & K(N_T, N_T) + R \end{bmatrix}\right). \tag{38}$$

The additional "kernel" notation $\Sigma_f \equiv K(N_T, N_T)$ is ubiquitous in GPR. Time-domain correlations specified by $R(\nu)$ populate each element of a matrix $K(\cdot,\cdot)$, where the dimensions of the matrix depend on the vector length of each argument. For example, for $K(N_T, N_T)$, the notation defines a square matrix where diagonals correspond to $\nu = 0$, and off-diagonal elements correspond to the separation of two arbitrary points in time, i.e., $\nu \neq 0$.

Following Ref. [41], the moments of the conditional predictive distribution $\Pr(f^\ddagger|y)$ can be derived from the joint distribution $\Pr(y, f^\ddagger)$ via standard Gaussian identities:

$$\mu_{f^\ddagger|y} = \mu_f + K(N^\ddagger, N_T)[K(N_T, N_T) + R]^{-1}(y - \mu_y), \tag{39}$$

$$\Sigma_{f^\ddagger|y} = K(N^\ddagger, N^\ddagger) - K(N^\ddagger, N_T)[K(N_T, N_T) + R]^{-1}K(N_T, N^\ddagger). \tag{40}$$

The prediction procedure outlined above holds true for any choice of kernel $R(\nu)$. In any GPR implementation, the data set, y, constrains the prior model, yielding an posterior predictive distribution. The mean values of this predictive distribution, $\mu_{f^\ddagger|y}$, are the state predictions for the qubit under dephasing at test points in $N^\ddagger$.

In our work, we focus on a "periodic kernel" (PER) to encode a covariance function which is theoretically guaranteed to approximate any zero-mean covariance-stationary process, f, in the mean-square limit, by having the same structure as a covariance function for trigonometric polynomials with infinite harmonic terms [38,42]. The sine-squared-exponential kernel represents an infinite basis of oscillators and is defined as

$$R(\nu) \equiv \sigma^2 \exp\left(-\frac{2\sin^2\left(\frac{\omega_0^{(B)}\nu}{2}\right)}{l^2}\right). \tag{41}$$

This kernel is described using just two key hyperparameters: the frequency-comb spacing for our infinite basis of oscillators, ω_0, and a dimensionless length scale, l. We use physical sampling considerations to approximate their initial conditions prior to an optimization procedure, namely, that the longest correlation length encoded in the data sets the frequency resolution of the comb, and the scale at which changes in f are resolved is limited physically by the minimum time taken between sequential Ramsey measurements:

$$\frac{\omega_0^{(B)}}{2\pi} \sim \frac{1}{\Delta t N}, \tag{42}$$

$$l \sim \Delta t. \tag{43}$$

Because the periodic kernel can be shown to be formally equivalent to the basis of oscillators employed in the LKFFB algorithm in a limiting case (see Appendix C for a discussion using the results in Ref. [42]), the inclusion of GPR using this kernel permits a comparison of the underlying algorithmic structures for the task of predictive estimation using spectral methods.

For the analysis of covariance-stationary time series under a GPR framework, we deemphasize popular kernel choices such as a RBF, a RQ, and Matern kernels (e.g., MAT32) [41,43]. An arbitrary-scale mixture of zero-mean Gaussian kernels probes an arbitrary area around zero in the Fourier domain, as schematically depicted in Fig. 2(a). While such kernels capture the continuity assumption that is ubiquitous in machine learning, they are structurally inappropriate for probing a process characterized by an arbitrary power spectral density (e.g., Ohmic noise). Another common kernel for time-series analysis is a quasiperiodic kernel (QPER) defined by a product of a RBF with a periodic kernel [44]. This product corresponds to a convolution in the Fourier domain giving rise to a comb of Gaussians at the expense of an increase in the number of parameters required for kernel tuning. One can also consider specific types of $\mathrm{AR}(q)$ processes using Matern kernels of order $q + 1/2$, but with increased restrictions on the form of the coefficients [41,45]. A simple consideration of autoregressive approaches suggests that a Matern kernel for $q = 1$ (MAT32) can be briefly trialed under GPR, whereas high-q autoregressive processes are naturally and generally treated under a KF framework. Further discussion of kernel choice appears in Sec. V.

IV. ALGORITHM PERFORMANCE CHARACTERIZATION

In the results to follow, our metric for characterizing performance of optimally tuned algorithms is the normalized Bayes prediction risk:

$$\tilde{L}_{\mathrm{BR}} \equiv \frac{L_{\mathrm{BR}}(n|I)}{\langle (f_n - \mu_f)^2 \rangle_{f,\mathcal{D}}}, \qquad \mu_f \equiv 0. \tag{44}$$

A desirable forward prediction horizon corresponds to maximal $n^* \in [0, N_P]$ for which the normalized Bayes prediction risk at all time steps $n \leq n^*$ is less than unity. We compare the difference in maximal forward prediction horizons between algorithms in the context of realistic operating scenarios. We begin here by introducing the numerical methods employed for generating data sets on which predictive estimation is performed.

We simulate environmental dephasing through a Fourier-domain procedure described in Appendix A 2 [46] in order to simulate an f which is mean-square ergodic and covariance stationary. For the results in this paper, we choose a flattop spectrum with a sharp high-frequency cutoff for simplicity, as this choice of power spectral density theoretically favors no particular choice of algorithm but violates the Markov property.

In our simulations, we also must mimic a measurement process which samples the underlying "true" dephasing process. The algorithmic parameters $\{N_T, \Delta t\}$ represent a sampling rate and Fourier resolution set by the simulated measurement protocol; we choose regimes where the Nyquist rate $r \gg 2$. In generating noisy simulated measurement records, we corrupt a noiseless measurement by additive Gaussian white noise. Since f is Gaussian, the measurement noise level, NL, is defined as the ratio between the standard deviation of additive Gaussian measurement noise, $\sqrt{R}$, and the maximal spread of random variables in any realization f. We approximate the maximal spread of f as three sample standard deviations of one realization of a true f value, $\mathrm{NL} = \sqrt{R}/3\sqrt{\hat{\Sigma}_f^{n,n}}$. The use of a hat in this notation denotes sample statistics. This computational procedure enables a consistent application of measurement noise for f from arbitrary, non-Markovian power spectral densities. For the case where binary outcomes are required, we apply a biased coin flip using Eq. (34).

A. Algorithmic optimization

All algorithms in this paper employ machine-learning principles to tune unknown design parameters based on training data sets. The physical intuition associated with optimizing our filters is that we are cycling through a large class of general models for environmental dephasing and seeking the model(s) which best fit the data, subject to various constraints. Optimizing over a general class of models allows each filter to track stochastic qubit dynamics under arbitrary covariance-stationary, non-Markovian dephasing. We elect to deploy an optimization routine with minimal computational complexity to enable nimble deployment of KF and GPR algorithms in realistic laboratory settings, particularly since LSF optimization is extremely rapid for our application [28].

Kalman filtering in our setting poses a significant challenge for general optimizers, as the lack of theoretical bounds on the values of (σ, R) results in large, flat regions of the Bayes-risk function. Furthermore, the recursive structure of the Kalman filter means that no analytical gradients are accessible for optimizing a choice of cost function, and a large computational burden is incurred for any optimization procedure. We randomly distribute (σ_k, R_k) pairs for $k = 1, \ldots, K$ over 10 orders of magnitude in two dimensions in order to sample the optimization space.

We then generate a sequence of loss values $L(\sigma_k, R_k)$ for each k value by considering a small region around $n = 0$, where the size of the region is an n_L number of time steps and we look forward or backward from $n = 0$:

$$L(\sigma_k, R_k) \equiv \sum_{n=1}^{n_L} L_{\mathrm{BR}}(n|I = \{\sigma_k, R_k\}). \tag{45}$$

Here, $L_{\mathrm{BR}}(n|I = \{\sigma_k, R_k\})$ is given by Eq. (2), and it is summed over $0 \leq n_L \leq |N_T|$ $(0 \leq n_L \leq |N_P|)$ backward (forward) time steps for state estimation (prediction). In the notation for I above, we omit Kalman dynamical model design parameters for an ease of reading. Typically, I would include, for instance, the set of autoregressive coefficients in AKF and the set of fixed basis frequencies in LKFFB. Values of n_L are chosen such that the sequence $\{L(\sigma_k, R_k)\}$ defines sensible shapes of the total loss function over parameter space and the numerical experiments in this paper. A choice of small n_L in state estimation ensures that data near the prediction horizon are employed—a region where the Kalman filter is most likely to converge. Similarly, in state prediction, large n_L will flatten the true prediction loss function, as long-term prediction errors dominate smaller loss values occurring during the short-term prediction period. In addition, one can weight state estimation and state prediction loss functions differently by choosing different values of n_L for state estimation and prediction, though we set n_L to be the same in both regions. While simple and by no means optimal, our tuning approach is computationally tractable and efficient compared to the application of standard optimization routines, where each loss value calculation requires a recursive filter to act on a long measurement record. Furthermore, our approach ensures that tuning procedures are performed offline such that a tuned algorithm is simple in its recursive structure and performs rapid calculations at each time step.

An ideal parameter pair (σ^*, R^*) minimizes the Bayes risk over K trials for both state estimation and prediction. We define acceptable low-loss regions for state estimation and prediction as being the set which returns a loss that is less than 10% of the median risk over K trials. In the event that low-risk regions do not exist for both state estimation and prediction for a given parameter pair, we deem the optimization to have failed, as the state estimation performance is uncorrelated with the forward prediction [for an illustration, see panel Fig. 7(h)].

In GPR, the set of parameters $I = \{\sigma, R, \omega_0^{(B)}, l\}$ requires optimization. However, in contrast to the KF, no recursion exists and analytic gradients are accessible to simplify the overall optimization problem. Instead of minimizing Bayes state estimation risk, we follow a popular practice of maximizing the Bayesian likelihood. Initial conditions and optimization constraints are derived from physical arguments as described in Sec. III.

B. Performance of the KF using linear measurement

The general performance of the various KF algorithms discussed above is illustrated in Fig. 4, which compares the AKF and LKFFB algorithms using a linear measurement record. Here, the solid black line represents the underlying true f, and solid markers indicate noisy simulated linear measurement data. Future predictions using the various KF formalisms and the (nonrecursive) LSF filter [28] are shown as colored open markers, based on these data. The selected single realization of the prediction process demonstrated in Fig. 4(a) is representative of a broad ensemble of simulated data sets and demonstrates the ability of all algorithms to perform a future prediction with varying degrees of success.

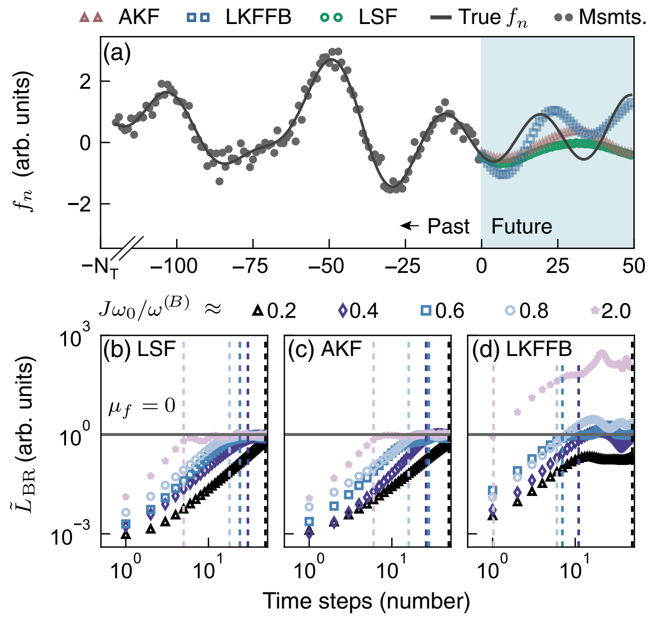

FIG. 4. (a) Solid dots depict y_n against time steps n, and data collection ceases at $n = 0$. Optimized LSF, AKF, and LKFFB yield predictions $n > 0$ in the blue region plotted as open, colored markers. A black solid line shows one realization of a true f_n, drawn from a flattop spectrum with J true Fourier components spaced ω_0 apart and uniformly randomized phases. Other parameters are $\omega_0/\omega_0^{(B)} \notin \mathcal{Z}$ (natural numbers); $J = 45\,000$; $\omega_0/2\pi = \frac{8}{9} \times 10^{-3}$ Hz, such that > 500 true components fall between adjacent LKFFB oscillators; and NL = 10%. (b)–(d) The procedure in (a) is repeated for ensemble M different realizations of f and noisy data sets to compute $\tilde{L}_{\rm BR}$ for a LSF, an AKF, and a LKFFB. $\tilde{L}_{\rm BR}$ vs $n \in [0, N_P]$ is plotted; the dark-gray horizontal line marks $\tilde{L}_{\rm BR} \equiv 1$ for predicting the mean $\mu_f \equiv 0$ value. Vertical dashed lines mark the forward prediction horizon, n^*, where $\tilde{L}_{\rm BR} \lesssim 0.8 < 1$ for all prediction time steps $0 < n \leq n^*$ in outperforming the prediction of the noise mean. Marker color (dark indigo to pink) depicts the true f cutoff, $J\omega_0$, varied relative to $\omega^{(B)} \equiv \omega_0^{(B)} J^{(B)} \approx r\omega_{(S)}$, with a fixed Nyquist $r \gg 2$, $\omega_0/2\pi = 0.497$ Hz, $J = 20$, 40, 60, 80, and 200; NL = 1%. In (a)–(d), a trained LKFFB is implemented with $\omega_0^{(B)}/2\pi = 0.5$ Hz and $J^{(B)} = 100$ oscillators; trained AKF and LSF models are $q = 100$, with $N_T = 2000$, $N_P = 50$ steps, $\Delta t = 0.001$ s, $M = 50$ runs, and $K = 75$ optimization trials.

In general, our objective is to maximize the forward prediction horizon, n^*, in any algorithmic setting. In Figs. 4(b)–4(d), we explore the key determining factors, setting the value of the prediction horizon under the three main Kalman filtering algorithms treated here. We plot the ensemble-averaged $\tilde{L}_{\rm BR}$ as a function of forward prediction time when adjusting the ratio of the cutoff frequency in the noise, $J\omega_0$, to the sample rate in the measurement routine ($\omega_{(S)} = 2\pi/\Delta t$) without physical aliasing, such that the Nyquist $r \gg 2$ and $\omega_{(S)} \approx \omega^{(B)}/r$, where $\omega^{(B)}$ incorporates a (potentially incorrect) bandwidth assumption about dephasing noise for the LKFFB. Here again, we have a forward prediction horizon for time steps $0 < n < n^*$ if $\tilde{L}_{\rm BR} \lesssim 1$ for all time steps in this region, and an algorithm seeks to maximize n^*. In this region, each algorithm predicts future dynamics better than naively predicting the mean behavior of f ($\mu_f \equiv 0$), indicated by a dark-gray horizontal line.

The prediction horizon, indicated approximately by dashed vertical lines, for all algorithms increases as the measurement becomes sufficiently fast to sample the highest frequency dynamics of f. We confirm numerically that the absolute prediction horizons for any algorithm are arbitrary and adjustable through the sample rate, allowing us to restrict our analysis to comparative statements between algorithms for future results. While differences between protocols appear to be reasonably small, we note that, in most cases examined, the AKF demonstrates superior performance to the LKFFB, subject to the realistic constraint that the true dynamics of f cannot be perfectly projected onto the basis used in the LKFFB (the latter situation corresponds to substantial prior knowledge of the dynamics of f). The role of undersampling in the LKFFB becomes pronounced as predictive estimates lead to unstable behavior relative to the naive prediction of $\mu_f = 0$ in the case $J\omega_0/\omega^{(B)} = 2$ in Fig. 4(d). The AKF and the LSF share autoregressive coefficients, and therefore both algorithms demonstrate comparable $\tilde{L}_{\rm BR}$ prediction risk in the ensemble average.

A key implied benefit of the use of Kalman filtering vs the LSF with high-order autoregressive dynamics alone is the addition of robustness against measurement noise. In order to probe measurement noise filtering capability numerically, we perform direct comparisons of filter performance under varying measurement-noise strength for both the AKF and the LSF. Since autoregressive coefficients learned in (noisy) environments are recast in Kalman form, we test measurement-noise filtering in Kalman frameworks enabled by the design parameter R. In Fig. 5(a), we plot the $\tilde{L}_{\rm BR}$ prediction risk for the AKF and the LSF as a ratio such that a value greater than unity implies that the LSF outperforms the AKF. In cases (i)–(iv), we increase the applied noise level to our data sets $\{y_n\}$ representing simulated measurements on f. For the applied measurement NL > 1% in (ii)–(iv), we find that the AKF or LSF < 1 and the AKF outperforms the LSF for the conditions studied here, with a general trend towards increasing benefits as noise increases, until the noise becomes so large (iv) that the benefits fluctuate as a function of n. Calculations of the ensemble-averaged $\tilde{L}_{\rm BR}$ in Fig. 5(b) demonstrate that all ratios reported in Fig. 5(a) correspond to a useful forward prediction horizon.

In machine-learning or optimal control settings, the robustness of the learning procedure to small changes in the underlying system is an essential characteristic of the algorithm. In our case, we have already seen that the quality of projection of the true dynamics of f onto the LKFFB basis can have a significant impact on the quality of learning and the predictive estimation. We now explore this initial finding in more detail.

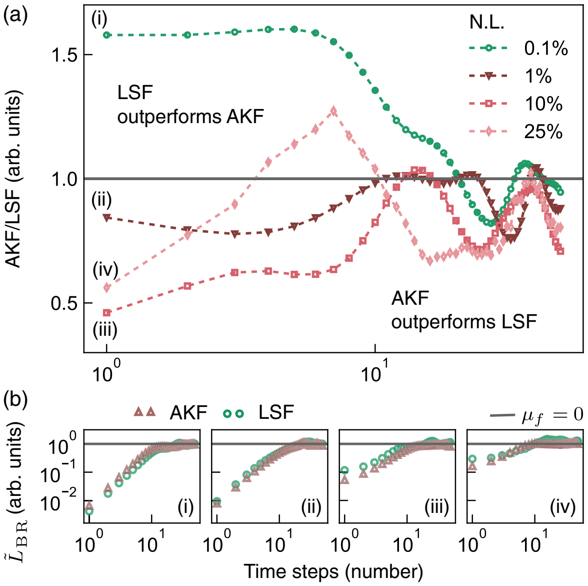

FIG. 5. Measurement noise filtering in AKF vs LSF. (a) Dashed lines with markers depict the ratio of $\tilde{L}_{\rm BR}$ for AKF to LSF against time steps $n > 0$, for cases (i)–(iv) with NL = 0.1%, 1.0%, 10.0%, and 25.0%. The green trajectory shows that the LSF outperforms the AKF with a ratio > 1 for $n \leq n^*$; crimson trajectories show that the AKF outperforms the LSF with a ratio < 1 for $n \leq n^*$. (b) $\tilde{L}_{\rm BR}$ against n is plotted for cases (i)–(iv), which confirms that a maximal forward prediction horizon marked by n^* exists for all ratios in (a) for both the LSF and the AKF. In (a) and (b), the AKF and the LSF share identical $\{\phi_q\}$. True f is drawn from a flattop spectrum with $\omega_0/2\pi = \frac{8}{9} \times 10^{-3}$ Hz, $J = 45\,000$, $N_T = 2000$, $N_P = 100$ steps, $\Delta t = 0.001$ s, and $r = 20$, such that Fig. 6(c) corresponds to case (ii). The AKF is optimized with $q = 100$, $M = 50$ runs, and $K = 75$ trials.

In Fig. 6, we simulate various learning conditions, including (a) perfect learning in the LKFFB, (b) imperfect projection relative to the LKFFB basis, (c) imperfect projection combined with finite algorithm resolution, and (d) imperfect learning and undersampling relative to the true noise bandwidth. The ordering of the figure presentation highlights the degree of impact of the introduced pathologies on the LKFFB. By contrast, we find reasonable model robustness in the AKF and the LSF at the expense of performance in the somewhat unrealistic perfect learning case.

We expose the underlying optimization results for choosing an optimal (σ^*, R^*) for the LKFFB in Figs. 6(e)–6(h) and for the AKF in Figs. 6(i)–6(l). Individual sample points are highlighted as solid dots, while low-loss pairs in this 2D space are highlighted for giving low state estimation (purple) or prediction (crimson) risk via the shaded circles. As the model pathologies indicated above increase, these data demonstrate a divergence between regions of the

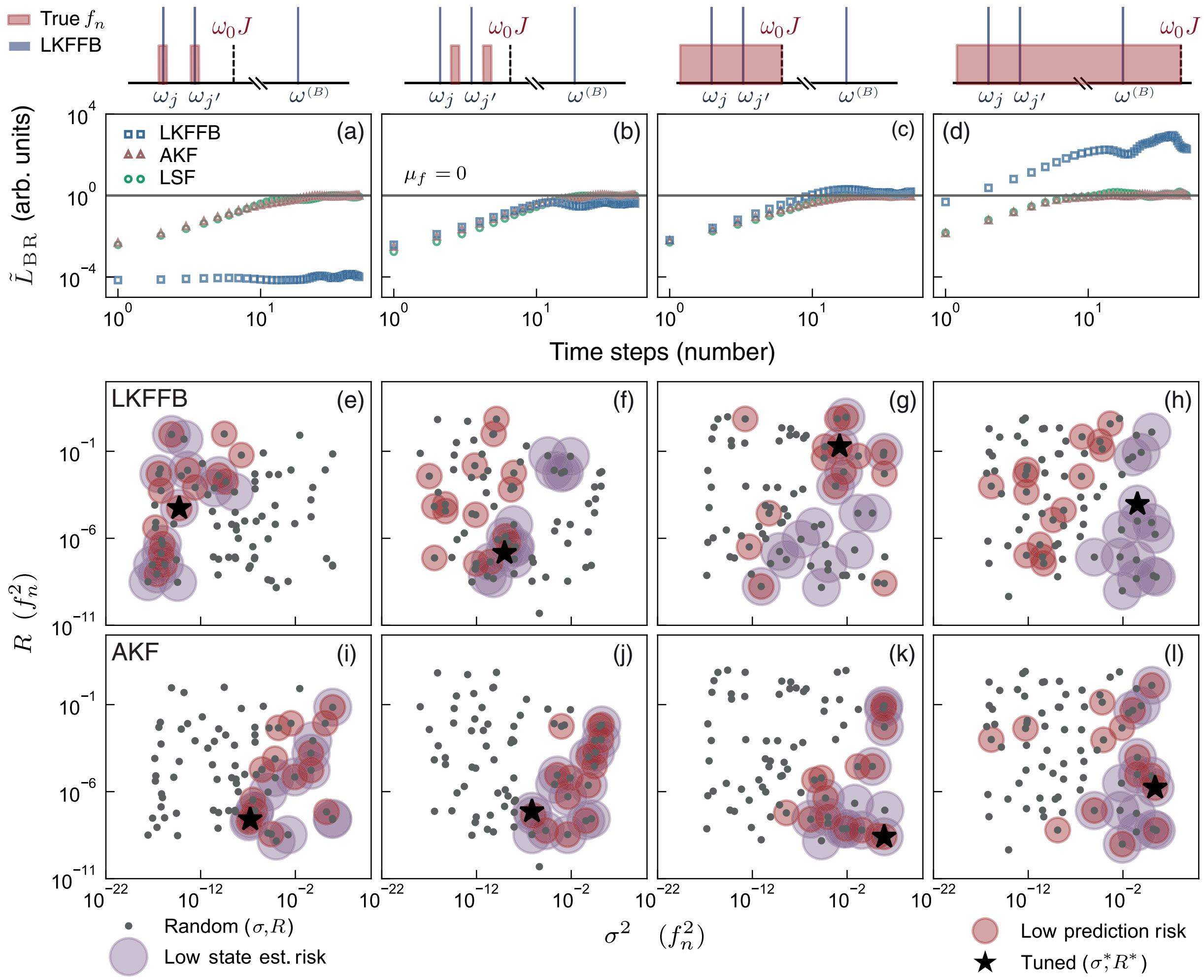

FIG. 6. Comparison of KF performance under various imperfect learning scenarios. (a)–(d) True noise properties are varied to introduce pathological learning with respect to a fixed algorithmic configuration: $\omega_0/2\pi = 0.5$, 0.499, $\frac{8}{9} \times 10^{-3}$, $\frac{8}{9} \times 10^{-3}$ Hz, and $J = 80$, 80, 45 000, and 80 000 respectively. The relationship between the LKFFB basis and the true noise spectrum is shown schematically above the columns. (a) Perfect learning. (b) Imperfect projection on the LKFFB basis. (c) Finite computational Fourier resolution. (d) Relaxed basis bandwidth assumption. (a)–(d) $\tilde{L}_{\rm BR}$ against time steps $n > 0$ is shown for the LKFFB, the AKF, and the LSF. (e)–(l) Optimization results for (top row) the LKFFB and (bottom row) the AKF in each of the four regimes in (a)–(d). The gray dots depict K random (σ^2, R) pairs, where M realizations of f, $\mathcal{D}$ are used to calculate $\tilde{L}_{\rm BR}$ for each pair. Purple (crimson) circles represent low-loss regions where the risk value in Eq. (45) for (σ^2, R) is $< 10\%$ of the median risk value during state estimation (prediction) for $-n_L < n < 0$ $(n_L > n > 0)$, with $n_L = 50$. The black star, (σ^*, R^*), minimizes risk values over the purple circles during state estimation. A KF filter is "tuned" if an optimal (σ^*, R^*) value lies in the overlap of low-loss regions for state estimation (purple) and prediction (crimson); disjoint regions in (h) show LKFFB tuning failure. KF algorithms set up with $q = 100$ for the AKF; $J^{(B)} = 100$ and $\omega_0^{(B)}/2\pi = 0.5$ Hz for the LKFFB, with $N_T = 2000$, $N_P = 100$ steps, $\Delta t = 0.001$ s, and $r = 20$; and NL = 1%.

optimization space which permit low-loss state estimation and forward prediction for the LKFFB. By contrast, the overlap in low-loss Bayes-risk regions does not change for the AKF across Figs. 6(i)–6(l).

The Kalman filtering algorithms employed here combine recursive state estimation with the establishment of a dynamical model in the Fourier domain. Therefore, one way to explore algorithmic performance is to look directly at the efficacy of spectral estimation relative to the true (here, numerically engineered) hidden dynamics of f. For both the LKFFB and the AKF, we plot the extracted power spectral density, $S(\omega)$, as a function of the angular frequency, ω, for different measurement sampling conditions in Fig. 7 against the true spectrum used to define f. These simulated experimental conditions match those introduced in Fig. 4(b).

In the case of the LKFFB, we plot the learned instantaneous amplitudes from a single run (blue markers), and for the AKF, we extract the optimized algorithm parameters as described above (red markers). Under the assertion that the LSF implements an AR(q) process, the set of trained parameters, $\{\{\phi_{q' \leq q}\}, \sigma^2\}$, from the AKF allows us to

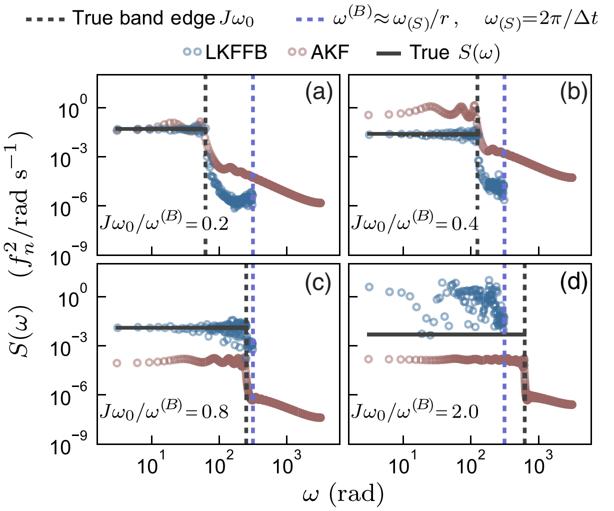

FIG. 7. (a)–(d) Blue (red) open markers plot the LKFFB (AKF) spectrum estimates, and the true spectrum (flattop) of f is plotted as a black solid line. The dashed black vertical line marks the true noise cutoff, $J\omega_0$, and this cutoff is varied relative to a measurement sampling rate, $\omega_{(S)}$, and $\omega^{(B)} \equiv \omega_0^{(B)} J^{(B)} \approx \omega_{(S)}/r$ in the LKFFB, such that $\omega_0/2\pi = 0.497$ Hz, $J = 20$, 40, 80, and 200. For the LKFFB, the blue open markers are $\propto \|\hat{x}_n^j\|^2$ in a single run with $\omega_0^{(B)}/2\pi = 0.5$ Hz for $j \in J^{(B)} = 100$ oscillators; the dashed blue vertical line marks the edge of the LKFFB basis. For the AKF, red markers are $\hat{S}(\omega)$ computed using the learned $\{\phi_{q' \leq q}\}$ and optimized σ^* values, with order $q = 100$. In all plots, the zeroth Fourier component is omitted on the log scale. $N_T = 2000$, $N_P = 50$ steps, $\Delta t = 0.001$ s, and $r = 20$, with $M = 50$ runs and $K = 75$ trials. NL = 1%.

derive experimentally measurable quantities, including the power spectral density of the dephasing process: $S(\omega) = \sigma^2 (2\pi |1 - \sum_{q'=1}^{q} \phi_{q'} e^{-i\omega q'}|^2)^{-1}$ [35].

The critical feature in these data sets is the existence of a flattop spectrum possessing a sharp high-frequency cutoff. Both classes of Kalman filtering algorithm successfully identify this structure and locate this high-frequency cutoff. In general, however, the LKFFB provides superior spectral estimation relative to the AKF and enables a better estimation of the signal strength in the Fourier domain, even in the presence of an imperfect projection of f onto the basis used in the LKFFB. The only case in which the LKFFB fails is in Fig. 7(d), where the LKFFB basis is ill specified relative to the true noise bandwidth. The observed behavior is somewhat surprising given the generally superior performance of the AKF in a predictive estimation, but it does highlight the practical difference between a Fourier-domain spectral estimation and a time-domain prediction.

C. Performance of the quantized Kalman filter

The discrete nature of projective measurement outcomes in quantum systems poses a potential challenge for Kalman filters in the event that measurement preprocessing as in Fig. 1(b) is not performed. We test filter performance for a predictive estimation when only binary measurement outcomes are available via the QKF. To reiterate, the QKF estimates and tracks hidden information, f_n, using the Kalman true state x_n. In our construction, the associated probability for a projective qubit measurement outcome, $\propto z_n$, is not inferred or measured directly but given deterministically by Born's rule encoded in the nonlinear measurement model, $z_n = h(f_n)$. The measurement action is completed by performing a biased coin flip, where z_n determines the bias of the coin.

For the QKF, the normalized ensemble-averaged prediction risk, $\langle (z_n - \hat{z}_n)^2 \rangle_{f,\mathcal{D}} / \langle (z_n - \mu_z)^2 \rangle_{f,\mathcal{D}}$, is calculated with respect to z as the relevant quantity parametrizing the qubit-state evolution, instead of the stochastic underlying f. This quantity is labeled "Norm. risk" in Fig. 8 and we test whether $\langle (z_n - \hat{z}_n)^2 \rangle_{f,\mathcal{D}} / \langle (z_n - \mu_z)^2 \rangle_{f,\mathcal{D}} < 1$ for $0 < n < n^*$ can be achieved for numerical experiments considered previously in the linear regime. In particular, we generate

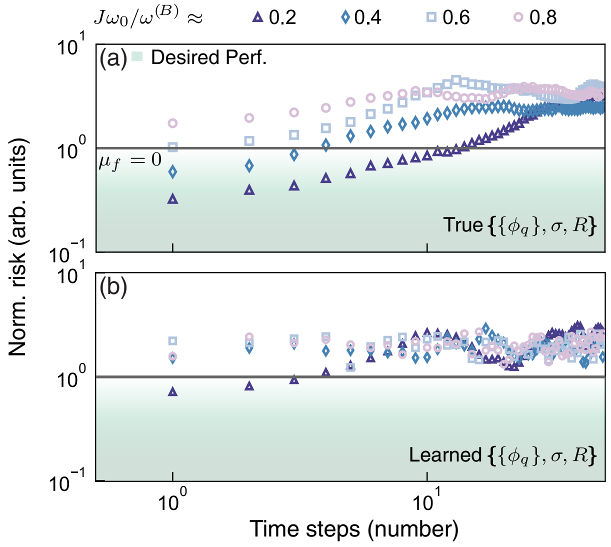

FIG. 8. Normalized risk against $n > 0$ plotted for a QKF in open markers; the dark-gray line at $\mu_f \equiv 0$ depicts performance underpredicting the noise mean. QKF outperforms predicting the mean if open markers lie in the green regions. Marker color (dark indigo to pink) depicts true noise cutoff varied by $J\omega_0/\omega_{(B)} = 0.2$, 0.4, 0.6, and 0.8 for f defined identically to Fig. 7 with $\omega_0/2\pi = 0.497$ Hz, $J = 20$, 40, 60, and 80; NL = 1%. (a) We obtain $\{\phi_{q' \leq q}\}$, $q = 100$ coefficients from AKF or LSF acting on a linear measurement record generated from true f. A new truth, f', is generated from an AR(q) process using $\{\phi_{q' \leq q}\}$, $q = 100$ as true coefficients and by defining a known, true σ. Quantized measurements from f' are obtained; data are corrupted by measurement noise of a true, known strength R. (b) We use $\{\phi_{q' \leq q}\}$, $q = 100$ coefficients from (a), but we generate quantized measurements from the original, true f. QKF noise design parameters are optimized for $(\sigma^*_{\text{AKF}} \leq \sigma_{\text{QKF}}, R^*_{\text{AKF}} \leq R_{\text{QKF}})$ with $M = 50$ runs and $K = 75$ trials. For (a) and (b), $N_T = 2000$, $N_P = 50$ steps, and $\Delta t = 0.001$ s, $r \gg 2$.

true f defined in numerical experiments in Fig. 4(b) (and Fig. 7) for $q = 100$ and varying sample rates.

We isolate the role of the measurement action by first inputting into the QKF a true dynamical model rather than a dynamical model learned as in the standard AKF. To specify true dynamics, we begin with a set of $\{\phi_{q' \leq q}\}$ and exactly derive a new f'. As a result, the full set of parameters relevant to the filter, $\{\{\phi_{q' \leq q}\}, \sigma, R\}$, are perfectly defined and known, and the filter simply acts on single-shot qubit measurements. These simulations reveal that, subject to the generic measurement oversampling conditions introduced above, the QKF is able to successfully enable predictive estimation. As in the linear case, the absolute forward prediction horizon is arbitrary relative to $\omega_0 J/\omega^{(B)}$ and, implicitly, an optimization over the choice of q for a finite data size, N_T, in our application.

Our simulations reveal that the QKF is considerably more sensitive to measurement noise, model errors, and the degree of undersampling than the linear model, as shown in Fig. 8(b). Here, the QKF incorporates a learned dynamical model from an AKF in the linear regime, and we tune (σ, R) for use in the QKF. In particular, we explore $\sigma \geq \sigma^*_{\mathrm{AKF}}$ to incorporate model errors as $\{\phi_{q' \leq q}\}$ are learned in the linear regime. We also incorporate increased measurement noise via $R \geq R^*_{\mathrm{AKF}}$, as QKF receives raw data that have not been preprocessed or low-pass filtered. The underlying optimization problems are well behaved for all cases in Fig. 8(b) (not shown). As the sampling rate is reduced, the QKF forward prediction horizon collapses rapidly; i.e., there is a $\langle (z_n - \hat{z}_n)^2 \rangle_{f,\mathcal{D}} / \langle (z_n - \mu_z)^2 \rangle_{f,\mathcal{D}} > 1$ prediction risk for all $n > 0$.

D. Failure of GPR in predictive estimation

Under a GPR framework, we test whether predictive performance can be improved by considering the entire measurement record (at once) and projecting this record on an infinite basis of oscillators summarized by a periodic kernel. We investigate several different types of GPR models for $M = 50$ realizations of f in the top panel of Fig. 9. For the results shown, we use a popular choice of a maximum-likelihood optimization procedure implemented via the L-BFGS algorithm in GPy [47].

We find that the underlying optimization procedure for training on our measurement records remains difficult despite our having access to an analytical calculation for the cost function. For all results in Figs. 9(a) and 9(b), we use significant manual tuning prior to deploying the automated procedures in GPy. Hence, we focus on using numerical results under GPR to illuminate structural implications of the choice of kernels in our application, rather than making comparative statements about kernel performance.

The results we assemble demonstrate that the implementation of GPR with a periodic kernel critically depends

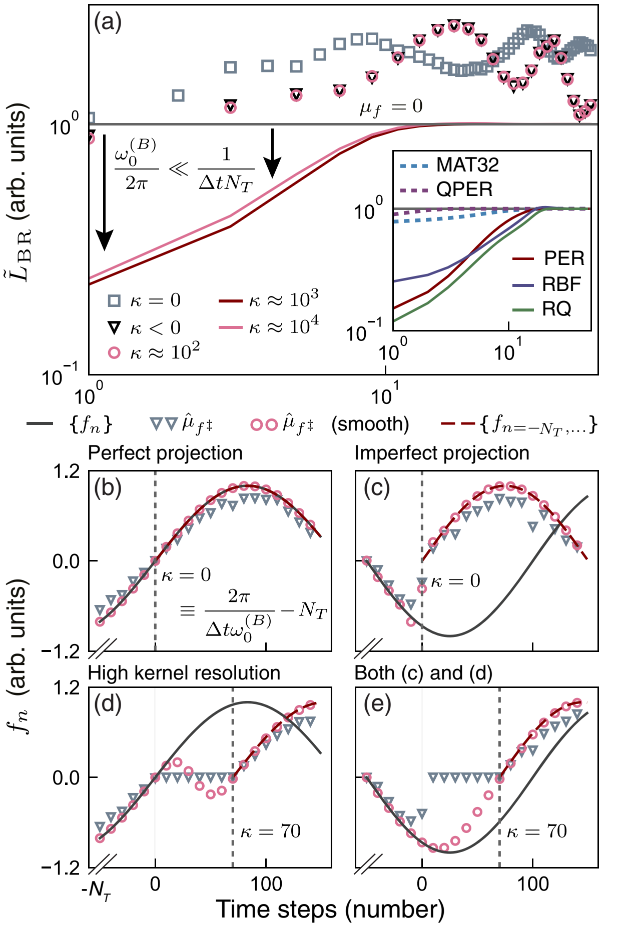

FIG. 9. (a) $\tilde{L}_{\mathrm{BR}}$ vs $n^\ddagger$ (in units of number of time steps) are plotted for GPR with a periodic kernel. The dark-gray horizontal line at unity for $\mu_f \equiv 0$ marks $\tilde{L}_{\mathrm{BR}}$ underpredicting the mean; GPR outperforms predicting the mean if data fall below this line. The gray-black markers correspond to optimization within physical bounds for $\kappa \leq 0$ (kernel resolution at or above Fourier resolution); crimson markers and lines depict optimization within unphysical regimes, $\kappa > 0$, with solid lines in the regime with high values of $\kappa \gg 0$. The remaining $\{R, \sigma, l\}$ values are optimized for non-negative values. (Inset) $\tilde{L}_{\mathrm{BR}}$ vs $n^\ddagger$ of a periodic kernel (PER) with $\kappa \approx 10^3$ is plotted against results from naively trained Gaussian kernels (RBF, RQ); a Matern kernel (MAT32) and a quasiperiodic kernel (QPER). (b)–(d) True state f_n vs n (the black solid line) and GPR predictions $\hat{\mu}_{f^\ddagger}$ vs $n^\ddagger$ (the open markers) plotted for a periodic kernel for tracking a sinusoid with frequency ω_0; the noisy data record (not shown) ceases at $n = 0$. We fix $\kappa = 0$ and 70; triangles plot predictions for manually tuned $\{R, \sigma, l\}$ values; circles plot predictions for the optimized $\{R, \sigma, l\}$ values. Vertical dashed lines mark $n = \kappa$, where we overlay true f at the beginning of the data record as a red dashed line. (b) Perfection projection is possible: $\omega_0/\omega_0^{(B)} \in \mathcal{Z}$ (natural numbers) and $\omega_0/2\pi = 3$ Hz. (c) Imperfect projection, with $\omega_0/\omega_0^{(B)} \notin \mathcal{Z}$, $\omega_0/2\pi = 3\frac{1}{3}$ Hz, and $\kappa = 0$. (d) We moderately raise $\kappa > 0$, such that $\omega_0/\omega_0^{(B)} \gg 0 \notin \mathcal{Z}$ for the original $\omega_0/2\pi = 3$ Hz. (e) We test (c) and (d) for $\kappa > 0$, $\omega_0/\omega_0^{(B)} \notin \mathcal{Z}$, and $\omega_0/2\pi = 3\frac{1}{3}$ Hz. For (b)–(e), $N_T = 2000$, $N_P = 150$ steps, and $\Delta t = 0.001$ s; NL = 1%.

on the frequency basis comb spacing, $\omega_0^{(B)}$, or, equivalently, a deterministic quantity, κ:

$$\kappa \equiv \frac{2\pi}{\Delta t \omega_0^{(B)}} - N_T. \tag{46}$$

The term $2\pi/\Delta t\omega_0^{(B)}$ is the theoretical number of measurements that, in principle, are required to *physically* achieve the Fourier resolution set by the kernel hyperparameter, $\omega_0^{(B)}$, and the fundamentally discrete nature of a sequential Ramsey measurement record, expressed by Δt. Hence, if $\kappa = 0$, the physical Fourier resolution determined by the data set matches the comb spacing in the periodic kernel. For $\kappa > 0$, the comb spacing in the periodic kernel is less than the Fourier spacing defined by the experimental data-collection protocol, with total measurements N_T.

In Fig. 9(a), we see that GPR predictive performance for the periodic kernel improves as the kernel's comb spacing is reduced. For each value of κ, we plot $\tilde{L}_{\mathrm{BR}}$ against time steps forward, $n^{\ddagger}$, where the double dagger corresponds to the evaluation of a predictive GPR distribution on arbitrarily chosen test points, $n^{\ddagger} = -N_T, \ldots, -1, 0, 1, \ldots, N_P$. Here, the optimizer is constrained to a region in $2\pi/\omega_0^{(B)}$ parameter space that corresponds to the order of magnitude for κ. The gray markers correspond to $\kappa \leq 0$, where the algorithm operates above (or at) the Fourier resolution. In this physically motivated parameter regime, the prediction fully fails. It is not until we set $\kappa \sim 10^3$—a nominally unphysical operating regime where the algorithm's frequency-comb spacing is smaller than the Fourier resolution—that the prediction succeeds (red traces). The latter case is physically difficult to interpret given that, in this regime, we find the best ensemble-averaged predictive performance only by providing unphysical freedom to the algorithm. We note that the optimized length scale for the periodic kernel remains on the order of $\Delta t \sim 10\Delta t$, such that, for all of the red trajectories in Fig. 9(a), we are operating in a high-$(2\pi/\omega_0^{(B)})$, low-l limit.

We contextualize the predictive performance of the GPR PER (the red solid line) in the high-κ, low-l limit by comparing it to predictions derived using other standard kernels (dotted lines) in the inset of Fig. 9(a). In such circumstances, the predictive performance of the periodic kernel prediction is on par with an application of a RBF and a scale mixture of zero-mean RQs. A Matern kernel (MAT32) and a QPER yield lower-than-anticipated performance. Further discussion of the choice of kernel appears in Sec. V. For each individual time trace contributing to the ensemble averages appearing here, we observe that all kernels (PERs, RBFs, RQs, MAT32s, and QPERs) yield good state estimations, and the state estimate at $n^{\ddagger} = -1$ agrees well with the truth. For GPR with PERs, RBFs, and RQs, the state estimate at $n^{\ddagger} = -1$ smoothly decays to the mean value (zero) for $n^{\ddagger} \geq 0$, and this effect yields a favorable normalized Bayes prediction risk immediately after $n^{\ddagger} > 0$, depicted by the solid lines in the inset of Fig. 9(a).

In order to illustrate the operating mechanism for the periodic kernel, we dramatically simplify the model used for f in Fig. 9(a) and replace it with a single-frequency sine curve. Figures 9(b)–9(e) demonstrate the prediction routine for GPR using a periodic kernel on a simplified version of f, and, as before, the predictions are always conducted from time step zero. For this simple example, the periodic kernel learns Fourier information in the measurement record enabling interpolation using test points $n^{\ddagger} \in [-N_T, 0]$ for the cases in all panels of Figs. 9(b)–9(e), and atypical features are seen only for test points in the prediction region. We consider predictions from a manually tuned model (the triangles) and an optimized GPR model where the remaining free $\{\sigma, R, l\}$ parameters are tuned using GPy (the circles).

An examination of different cases for imperfect learning reveal that this discontinuity exhibits deterministic behavior linked to the underlying structure of the algorithm, namely, to the value of κ. In our numerical experiments, we find that, in all cases, of imperfect learning under GPR with a periodic kernel, a discontinuity in the prediction sequence arises at $n^{\ddagger} = \kappa$. These discontinuities are marked by the vertical dashed lines in all panels of Figs. 9(b)–9(e). However, another feature appears which we identify as being linked to oversampling of the underlying process determining f. In such cases, the algorithm simply predicts zero out to $n^{\ddagger} = \kappa$ before discontinuously predicting future evolution, which does not appear to be similar to the true value of f. By contrast, an optimized model gives smoothly varying predictions which still adhere to the underlying behavior set by κ for $n^{\ddagger} > 0$.

In Figs. 9(b)–9(e), we also plot the value of f given from $n = -N_T$, the start of the data set, on top of the prediction from $n^{\ddagger} = \kappa$. Here, we see that the prediction provided by GPR matches the earliest stages of the underlying data set well. Through various numeric experiments, we find that the action of GPR in such parameter regimes (moderately positive values of $\kappa > 0$) appears to be to simply repeat the learned values of f from $n = -N_T$ beginning at $n^{\ddagger} = \kappa$. Accordingly, these predictions rarely describe the underlying forward dynamics of f well.

As we enter the high-κ regime, $\kappa \gg 0$, the features in Figs. 9(b)–9(e) disappear, and GPR predictions begin to track the (slow moving) "truth" when $n^{\ddagger} \gg 0$. Analogous to the inset in Fig. 9(a), we see the performance of PERs approach that of standard Gaussian kernels in this simplified case.

V. DISCUSSION

The numeric simulations we perform in this work probe a wide variety of operating conditions in order to explore the algorithmic pathologies of leading forecasting techniques drawn from engineering, econometrics, and

TABLE I. Overview of performance results for all algorithms in this work across all frameworks. Column 2 lists mechanisms for data input (recursive or batch) and the key structural comparisons being made between algorithms. Columns 3 and 4 qualitatively assess performance during qubit-state estimation ($n < 0$) and prediction ($n \geq 0$); we comment on the conditions in which algorithms are found to perform strongly or to fail in columns 5 and 6.

Algorithm	Structure	State Estimation	Prediction	Advantages	Weaknesses
Kalman, AKF	Recursive; autoregressive dynamical model	Good	Best	Robust to measurement noise and variety of operating regimes	Need to train AR model prior to filtering and prediction
Kalman, LKFFB	Recursive; Fourier synthesized dynamical model	Good	Moderate	Robust to measurement noise	Oscillator structure not robust in all operating regimes
Kalman, QKF	Recursive; single-qubit data, autoregressive dynamical model	Moderate	Moderate	Direct processing of single-shot qubit data	Susceptible to rapid error accumulation via model nonlinearities and binary data
Least squares, LSF	Batch processing; linear regression	Good	Good	Rapid extraction of autoregressive dynamics from large data sets	Not robust against measurement noise
GPR (PER)	Batch processing; Bayesian data constrained model selection	Good	Poor	Good pattern interpolation during state estimation	Susceptible to producing numeric artifacts in forward prediction

machine-learning communities when applied to the predictive estimation of qubit evolution. A qualitative summary of our observations and key algorithmic differences is given in Table I for ease of reference.

Our central finding is that, overall, the autoregressive Kalman filter provides an effective path to perform both state estimation and forward prediction for non-Markovian qubit dynamics. Recasting dynamics into an AKF filter, importantly, provides model robustness against details of the underlying dynamics as well as a filtering of noise that allows it to outperform the simpler LSF in Ref. [28]. Measurement noise filtering is enabled in the Kalman framework through the optimization procedure for R and has a regularizing (smoothing) effect. Additionally, optimization of the imperfectly learned dynamical model is provided through the tuning of σ. The joint optimization procedure over (σ, R) ensures that the relative strength of the noise parameters is also optimized.

The AKF is also demonstrated to work well with discretized projective measurement models via what we refer to as the QKF. In the QKF, we employ single-shot, discretized qubit data while enabling model-robust qubit-state tracking and increased measurement-noise filtering via the underlying AKF algorithm. However, we find that the QKF is vulnerable to the buildup of errors for arbitrary applications, and we provide three explanatory remarks from a theoretical perspective. First, the Kalman gains are recursively calculated using a set of linear equations of motion which incorporate the Jacobian H_n of $h(x_n)$ at each n. All nonlinear Kalman filters perform well if errors during filtering remain small such that the linearization assumption holds at all time steps. Second, measurements are quantized, and hence residuals must be $\{-1, 0, 1\}$ rather than continuously represented floating-point numbers. In our case, the Kalman update to x_n at n mediated by the Kalman gain cannot benefit from a gradual reduction in residuals. A third effect incorporates the consequences of both quantized residuals and a nonlinear measurement action. In linear Kalman filtering, Kalman gains can be precalculated in advance of the acquisition of any measurement data: the recursion of Kalman state variances P_n can be decoupled from the recursion of Kalman state means, x_n [31]. In our application, quantized residuals affect the Kalman update of x_n and, furthermore, they affect the recursion for the Kalman gain via the state-dependent Jacobian, H_n.

In this context, we demonstrate numerically that the QKF achieves a desirable forward prediction horizon when the buildup of errors during filtering is minimized, for example, by specifying Kalman state dynamics and noise strengths perfectly, and/or by severely oversampling relative to the true dynamics of f. At present, we simply interpret our results on the QKF as a demonstration that one may, in principle, track stochastic qubit dynamics using single-shot measurements under a Kalman framework. The QKF also has the benefit, as constructed, of reverting to the AKF if suitable preprocessing of data is performed prior to execution of the iterative state estimation algorithm. In common laboratory settings, the measurement protocol may be effectively linearized through a simple averaging of multiple single-shot measurements, the application of Bayesian estimation protocols, or other preprocessing identified above. So long as the preprocessing takes place on timescales that are fast relative to the underlying qubit dynamics, the measurement linearization has no impact other than to change the effective sample rate of the measurements. Thus, it is our view that full implementation of the QKF is not essential if improved optimization routines are not accessible.

It is possible that QKF forward prediction horizons in realistic learning environments can be improved by solving the full $q+2$ optimization problem for $\{\{\phi_{q'\leq q}\}, \sigma, R\}$, rather than by employing the approach taken in this paper. However, full optimization poses its own challenges given the observations we make about the optimization landscape, even for the 2D optimization problem faced in the AKF. More sophisticated, data-driven model-selection schemes are described for both KF and kernel learning machines (such as GPR) in the literature (see, e.g., Refs. [48,49]). Beyond standard local-gradient and simplex optimizers, we consider coordinate ascent [50] and particle swarm optimization techniques [51] to be promising, nascent candidates, and their application remains an open research question. One may also consider switching from a high-order AR(q) to an ARMA model with a smaller number of optimization parameters. Typically, this switch is accomplished by incorporating greater prior information about the underlying dynamic process in the design of the ARMA model and/or using model-less particle-based or unscented filtering techniques to overcome nonlinearities in an ARMA representation (see, e.g., Ref. [2]). The latter set of techniques are well adapted for nonlinear models but are likely to require a modification to allow for non-Markovian dynamics (e.g., by designing an appropriate transition probability for otherwise Markov resampling procedures); by contrast, a typical recursive ARMA formulation for our application may track temporal correlations but be ill equipped for nonlinear coin-flip measurements. One expects a straightforward application of such procedures to be complicated.

Our general results on the use of autoregressive models for building Kalman dynamical models stand in contrast to Fourier-domain approaches in the LKFFB and GPR using a periodic kernel; both show significant performance degradation in cases when the learning of state dynamics is imperfect. In investigating the loss of performance for the LKFFB, we find that the efficacy of this approach depends on a careful choice of a *probe* (i.e., a fixed computational basis) for the dynamics of f capturing the effect of dephasing noise on the qubit. In the imperfect learning regime of Fig. 4 and, identically, Fig. 7, the LKFFB reconstructs Fourier-domain information to high fidelity across a range of sampling regimes but is outperformed by the AKF in the time domain (Fig. 4). Since the LKFFB tracks instantaneous amplitude and phase information explicitly for each basis frequency, the loss of the LKFFB time-domain predictive performance must accrue from a difficulty in tracking the instantaneous phase—rather than the amplitude—information.

While the difficulty of an instantaneous phase estimation is likely to be a disadvantage for the time-domain predictive performance of LKFFB, our results show that a Fourier-domain approach yields high-fidelity reconstructions of a power spectral density describing f. These reconstructions appear to be robust against imperfect projection on the LKFFB oscillator basis even as oversampling is reduced. These results suggest that an application of the LKFFB outside of predictive estimation could be tested against standard spectral estimation techniques in future work.

The challenge in adapting GPR for the task of a time-domain predictive estimation proves to be more striking. In our numerical simulations, under conditions comparable to those tested in the AKF, the values of a normalized Bayes prediction risk for all GPR models are at least an order of magnitude greater than the comparable performance of the AKF or LKFFB [refer to Fig. 5(b)(ii) and, equivalently, Fig. 6(c)]. This difference is somewhat surprising because, in the limit in which Γ_n is set to the identity in the LKFFB and an infinite basis of oscillators in the periodic kernel is truncated at the finite value, $J^{(B)}$, both the LKFFB and the GPR PER are formally equivalent to classical Kalman filtering for a collection of $J^{(B)}$ independent state-space resonators [42]. In this limit, the true f is described by theoretically identical covariance functions in both the KF and GPR frameworks. While we do not operate in this regime, one would expect the predictive capabilities of these two algorithms to be comparable.

In contrast to our observations for the various flavors of KF tested here, we observe that GPR predictions with a periodic kernel are useful for filtering or retrodiction but appear to have limited meaning for forward predictions for time steps $n = n^{\ddagger} > 0$. In our application, predictive performance of GPR with a periodic kernel for $\kappa = 0$ is shown to yield poor predictive performance over the ensemble average [Fig. 9(a)]. For the unexpected regime of $\kappa \gg 0$ and relatively small fixed l values, predictive performance improves and the periodic kernel performs similarly to RBFs and RQs. In this, a high-κ and a low-l regime, the sin term of the periodic kernel is slowly moving $[\sin(x) \approx x]$, and hence the argument of the exponential in the periodic kernel approximates a Gaussian, reducing to a RBF kernel. Our numerical investigations show that an optimized RQ kernel consistently chooses parameter regimes where a RQ also converges to a RBF. For the operating regimes pertinent to our application, it appears that the choice of the periodic, RBF, and RQ kernels produce theoretically equivalent results for forward predictions of the qubit state. In our analysis, these "forward predictions" simply arise from a smoothed decay of state estimates starting from test point $n^{\ddagger} = -1$ to the noise mean for test points $n^{\ddagger} > 0$, and they are difficult to interpret compared to their Kalman counterparts.

Our numerical characterization of the periodic kernel for a simple, noiseless f demonstrates that this kernel learns Fourier-domain amplitude information in a way that is better suited for pattern fitting than forward prediction. The predictive time-domain sequence of state estimates is repetitive at $n = n^{\ddagger} = \kappa$ and can be interpreted as successful qubit-state predictions only when f is perfectly learned (no discontinuities appear). When learning is imperfect, however, GPR with a periodic kernel is able to learn Fourier amplitudes to provide good retrodictive state estimates for $n^{\ddagger} < 0$, but forward predictions for $n^{\ddagger} > 0$ typically fail. Unlike the LKFFB, we believe the periodic kernel does not permit actively extracting and updating phase information for each individual basis oscillators at $n^{\ddagger} = \kappa$. Since phase information can be recast as amplitude information for any fixed-frequency oscillator, one would naively expect that forward predictions can be improved by increasing κ moderately, such that the higher-order terms in a series expansion of the sin term are nontrivial and $\sin(x) \approx x$ cannot apply. However, any positive value of κ means that we are probing dynamics at frequencies lower than those appearing in the data set. As such, a GPR-PER model predicts zero for $n^{\ddagger} \in [0, \kappa]$, $\kappa > 0$ before reviving at κ. The use of a procedure optimizing kernel noise parameters $\{\sigma, R\}$ does not change the behavior as $n^{\ddagger} \to \kappa$ but does smooth the discontinuities, as illustrated in Fig. 9(f). In letting $\kappa \gg 0$ (extremely large), we lose the uniqueness of the periodic kernel in summarizing an infinite basis of oscillators, and standard Gaussian kernels (e.g., RBF and RQ) are likely to apply.

It is possible that the choice of more-complex kernels could enhance forward time-series predictions via GPR, but they bring additional complications which currently remain unresolved in relation to the current application. As one example, our ability to use numerical investigations to inform kernel design is further distorted by the need for a robust optimization procedure, as illustrated by lower-than-anticipated predictive performance observed for QPERs. Another class of GPR methods—namely, spectral mixture kernels and sparse spectrum approximation using GPR—was explored in Refs. [52,53]. However, these techniques also require efficient optimization procedures to learn many unknown kernel parameters, whereas the sine-squared exponential in the periodic kernel is parametrized only by two hyperparameters. In addition to spectral methods, the generalization of MAT32 to higher $q + 1/2$ models probes only a subset of all possible AR(q) processes, as the restrictions on autoregressive coefficients in Matern kernels are greater than the general case considered under an AKF in this paper. A detailed investigation of the application of such methods for forward prediction beyond pattern recognition, and with limited computational resources, remains an area for future investigation.

VI. CONCLUSION

In this paper, we provide a detailed survey of machine-learning and filtering techniques applied to the problem of tracking the state of a qubit undergoing non-Markovian dephasing via a record of projective measurements. We specifically consider the task of performing predictive estimation: learning dynamics of the system from the measurement record and then predicting evolution forward in time. To accommodate stochastic dynamics under arbitrary dephasing, and without an *a priori* dynamical model, we choose two Bayesian learning protocols—GPR and KF. All Kalman algorithms predict the qubit state forward in time better than predicting mean qubit behavior, indicating successful prediction, though an autoregressive approach to building the Kalman dynamical model demonstrated enhanced robustness relative to Fourier-domain approaches. Forward prediction horizons could be arbitrarily increased for all Kalman algorithms by oversampling the underlying dephasing noise. Our investigations include studies of both linear and nonlinear measurement routines and validate the utility of the Kalman filtering framework for both. By contrast, under GPR, we find numerical evidence that this approach enables retrodiction but not forward predictions beyond the measurement record.

There are exciting opportunities for machine-learning algorithms to increase our understanding of dynamically evolving quantum systems in real time using projective measurements. Quantum systems coupled to classical spatially or temporally varying fields may benefit from classical algorithms to analyze correlation information and enable predictive control of qubits for applications in quantum information, sensing, and the like. Moving beyond a single qubit, we anticipate that measurement records will grow in complexity, allowing us to exploit the natural scalability offered by machine learning for mining large data sets. In realistic laboratory environments, the success of algorithmic approaches will be contingent on robust and computationally efficient algorithmic optimization procedures, as well as the extensions beyond Markovian dynamics studied here. The pursuit of these opportunities is the subject of ongoing research.

ACKNOWLEDGMENTS

The LSF filter is written by V. Frey and S. Mavadia [28]. The GPR framework is implemented and optimized using standard protocols in GPy [47]. The authors thank C. Granade, K. Das, V. Frey, S. Mavadia, H. Ball, C. Ferrie, and T. Scholten for the useful comments. This work is partially supported by ARC Centre of Excellence for Engineered Quantum Systems Project No. CE110001013, the U.S. Army Research Office under Contract No. W911NF-12-R-0012, and a private grant from H. & A. Harley.

APPENDIX A: PHYSICAL SETTING

In this appendix, we derive Eq. (1). We consider a qubit under environmental dephasing. For any two-level system, a quantum-mechanical description of physical quantities of interest can be provided in terms of the Pauli spin operators $\{\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\}$. If $\hbar\omega_A$ corresponds to an energy difference separating these two qubit states, then the Hamiltonian for a single qubit in free evolution can be written in the Pauli representation. We consider a qubit state in the $\hat{\sigma}_z$ basis, $|0\rangle$ or $|1\rangle$, with energies E_0 and E_1 in our notation, corresponding to a 0 or 1 outcome upon measurement. This physical setting yields a Hamiltonian for a single qubit as

$$\hat{\sigma}_z \equiv |1\rangle\langle 1| - |0\rangle\langle 0|, \tag{A1}$$

$$E_{0,1} \equiv \mp \frac{1}{2}\hbar\omega_A, \tag{A2}$$

$$\hat{\mathcal{H}}_0 = \frac{1}{2}\hbar\omega_A\hat{\sigma}_z. \tag{A3}$$

In this representation, the effect of dephasing noise on a free qubit system is that any initially prepared qubit superposition of $|0\rangle$ and $|1\rangle$ states will decohere over time in the presence of dephasing noise. This physical effect is modeled as a stochastically fluctuating process $\delta\omega(t)$ that couples with the $\hat{\sigma}_z$ operator. The noise Hamiltonian is described as

$$\hat{\mathcal{H}}_N(t) \equiv \frac{\hbar}{2}\delta\omega(t)\hat{\sigma}_z. \tag{A4}$$

In the formula above, $\delta\omega(t)$ is a classical, stochastically fluctuating parameter that models environmental dephasing, and $\hbar/2$ appears as a convenient scaling factor. The total Hamiltonian for a single qubit under dephasing is

$$\hat{\mathcal{H}}(t) \equiv \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_N(t). \tag{A5}$$

Since $\hat{\mathcal{H}}_N(t)$ commutes with $\hat{\mathcal{H}}_0$, we can transform away $\hat{\mathcal{H}}_0$ by moving to a rotating frame with respect to H_0. Let $|\psi(t)\rangle$ be a state in the lab frame, let $\hat{U}$ define a transformation to a rotating frame, and let $|\tilde{\psi}(t)\rangle$ be the state in the rotating frame. The tilde indicates operators and states in the transformed frame. In this simple case, the transformed Hamiltonian governing the evolution of $|\tilde{\psi}(t)\rangle$ is just $\hat{\mathcal{H}}_N(t)$:

$$\hat{U} \equiv e^{-i\hat{\mathcal{H}}_0 t/\hbar}, \tag{A6}$$

$$|\tilde{\psi}(t)\rangle \equiv \hat{U}^\dagger|\psi(t)\rangle, \tag{A7}$$

$$i\hbar\frac{d}{dt}|\tilde{\psi}(t)\rangle \equiv i\hbar\frac{d}{dt}\hat{U}^\dagger|\psi(t)\rangle \tag{A8}$$

$$= -\hat{\mathcal{H}}_0\hat{U}^\dagger|\psi(t)\rangle + i\hbar\hat{U}^\dagger\frac{d}{dt}|\psi(t)\rangle \tag{A9}$$

$$= (\hat{U}^\dagger\mathcal{H}(t)\hat{U} - \hat{\mathcal{H}}_0)|\tilde{\psi}(t)\rangle, \tag{A10}$$

$$\Rightarrow \hat{\tilde{\mathcal{H}}} \equiv \hat{U}^\dagger\mathcal{H}(t)\hat{U} - \hat{\mathcal{H}}_0 \tag{A11}$$

$$= \hat{U}^\dagger\hat{\mathcal{H}}_0\hat{U} + \hat{U}^\dagger\hat{\mathcal{H}}_N(t)\hat{U} - \hat{\mathcal{H}}_0 \tag{A12}$$

$$= \hat{\mathcal{H}}_N(t),\ [\hat{U}, \hat{\mathcal{H}}_0] = [\hat{U}, \hat{\mathcal{H}}_N(t)] = 0. \tag{A13}$$

In the semiclassical approximation, $\hat{\mathcal{H}}_N(t)$ commutes with itself at different t, and hence we can write a unitary time-evolution operator in the rotating frame as

$$\hat{\tilde{U}}(t, t+\tau) \equiv e^{-(i/\hbar)\int_t^{t+\tau}\hat{\mathcal{H}}_N(t')dt'} = e^{-(i/2)f(t,t+\tau)\hat{\sigma}_z}, \tag{A14}$$

$$f(t, t+\tau) \equiv \int_t^{t+\tau}\delta\omega(t')dt'. \tag{A15}$$

In the rotating frame, we prepare an initial state that is a superposition of $|0\rangle$ and $|1\rangle$ states. This state evolves under $\hat{\mathcal{H}}_N(t)$ during a Ramsey experiment for duration τ. Subsequently, the qubit state is rotated before a projective measurement is performed with respect to the $\hat{\sigma}_z$ axis; i.e., the measurement action resets the qubit.

Without loss of generality, define the initial state as $|\tilde{\psi}(0)\rangle \equiv (1/\sqrt{2})|0\rangle + (1/\sqrt{2})|1\rangle$ in the rotating frame. Then give the probability of measuring the same state after time τ in a single-shot measurement, d_n, as

$$\Pr(d_n = 1|f(0,\tau),\tau) = |\langle\tilde{\psi}(0)|\hat{\tilde{U}}(0,\tau)|\tilde{\psi}(0)\rangle|^2, \tag{A16}$$

$$\Pr(d_n = 0|f(0,\tau),\tau) \equiv 1 - \Pr(d_n = 1|f(0,\tau),\tau). \tag{A17}$$

The second $\pi/2$ control pulse rotates the state vector such that a measurement in the $\hat{\sigma}_z$ basis is possible, and the probabilities correspond to observing the qubit in the $|1\rangle$ state. Hence, Eq. (A16) defines the likelihood for a single-shot qubit measurement. Furthermore, Eq. (A16) defines the nonlinear measurement action on phase noise jitter, $f(0,\tau)$. We impose a condition that $f(0,\tau)/2 \leq \pi$, such that the accumulated phase over τ can be inferred from a projective measurement on the $\hat{\sigma}_z$ axis.

1. Experimentally controlled discretization of dephasing noise

In this section, we consider a sequence of Ramsey measurements. At time t, Eq. (A16) describes the qubit measurement likelihood at one instant under dephasing noise. We assume that the dephasing noise is slowly drifting with respect to a fast measurement action on timescales of order τ. In this regime, Eq. (A15) discretizes the continuous-time process $\delta\omega(t)$, at time t, for a number $n = 0, 1, \ldots, N$ equally spaced measurements, with $t = n\Delta t$. Performing the integral for $\tau \ll \Delta t$, we slowly drift the noise such that we substitute the following terms into Eq. (A15):

$$\delta\bar{\omega}_n \equiv \delta\omega(t')|_{t'=n\Delta t}, \tag{A18}$$

$$f_n \equiv f(n\Delta t, n\Delta t + \tau) \tag{A19}$$

$$= \frac{\hbar}{2}\int_{n\Delta t}^{n\Delta t+\tau} \delta\bar{\omega}_n dt' = \frac{\hbar}{2}\hat{\sigma}_z \delta\bar{\omega}_n \tau. \tag{A20}$$

In this notation, $\delta\bar{\omega}_n$ is a random variable realized at time $t = n\Delta t$, and it remains constant over a short duration of the measurement action, τ. We use the shorthand $f_n \equiv f(n\Delta t, n\Delta t + \tau)$ to label a sequence of stochastic, temporally correlated qubit phases $f \equiv \{f_n\}$.

Since the qubit is reset by each projective measurement at n, the unitary operator governing qubit evolution is also reset such that $\{\hat{\hat{U}}_n \equiv \hat{\hat{U}}(n\Delta t, n\Delta t + \tau)\}$ represents a collection of N unitary operators describing qubit evolution for each new Ramsey experiment. They are not to be interpreted, for example, as describing qubit free evolution without reinitializing the system. Hence, for each stochastic qubit phase f_n, the true probability for observing the $|1\rangle$ in a single shot is given by substituting f_n for $f(0,1)$ in Eq. (A16):

$$\Pr(d_n = d|f_n, \tau, n\Delta t) = \begin{cases} \cos\left(\frac{f_n}{2}\right)^2 & \text{for } d = 1 \\ \sin\left(\frac{f_n}{2}\right)^2 & \text{for } d = 0 \end{cases}. \tag{A21}$$

The last line follows from the fact that total probability of the qubit occupying either state must add to unity, yielding Eq. (1).

2. True dephasing noise engineering

In the absence of an *a priori* model for describing qubit dynamics under dephasing noise, we impose the following properties on a sequence of stochastic phases, $f \equiv \{f_n\}$, such that we can design meaningful predictors of qubit-state dynamics. We assert that a stochastic process, f_n, indexed by a set of values $n = 0, 1, \ldots, N$ satisfies

$$\mathbb{E}[f_n] = \mu_f \quad \forall n, \tag{A22}$$

$$\mathbb{E}[f_n^2] < \infty \quad \forall n, \tag{A23}$$

$$\mathbb{E}[(f_{n_1} - \mu_f)(f_{n_2} - \mu_f)] = R(\nu), \quad \nu = |n_1 - n_2|, \ \forall\, n_1, n_2 \in N, \tag{A24}$$

$$R(\nu) \neq \sigma^2\delta(\nu). \tag{A25}$$

Covariance stationarity of f is established by satisfying Eqs. (A22)–(A24), namely, that the mean is independent of n, the second moments are finite, and the covariance of any two stochastic phases at arbitrary time steps n_1, n_2, depend not on time steps but only on the separation distance, ν. The $\delta(\nu)$ in the last condition, Eq. (A25), is the Dirac-δ function and establishes that f is not δ correlated (white). This condition captures the slowly drifting assumption for environmental dephasing noise.

We also require that correlations in f eventually die off as $\nu \to \infty$; otherwise, any sample statistics inferred from noise-corrupted measurements are not theoretically guaranteed to converge to the true moments. Let M be the number of runs for an experiment with M different realizations of the random process f, μ_f be the true mean, $\hat{\mu}_f$ be its estimate, $\mathcal{D}_M$ denote the data set of M experiments, and $R(\nu)$ define the correlation function for the true process, f. Then mean-square ergodicity states that estimators approach true moments only if the correlations die off over long temporal separations:

$$\lim_{M\to\infty}\frac{1}{M}\sum_{\nu=0}^{M-1} R(\nu) = 0 \Leftrightarrow \lim_{M\to\infty}\mathbb{E}[(\hat{\mu}_f - \mu_f)^2]_{\mathcal{D}_M} = 0$$
$$\text{for } \nu = |n_{m_1} - n_{m_2}|,$$
$$\forall m_1, m_2 \in M, n_{m_1}, n_{m_2} \in N,$$
$$\text{with } \hat{\mu}_f = \frac{1}{M}\sum_{m=0}^{M} f_{n_m}. \tag{A26}$$

The statement above means that a true $R(\nu)$ value associated with f is bandlimited for sufficiently large (but unknown) values of M. If correlations never "die out," then any designed predictors for one realization of dephasing noise will fail for a different realization of the same true dephasing. For the purposes of experimental noise engineering, we satisfy the assumptions above by engineering discretized process, f, as

$$f_n = \alpha\omega_0 \sum_{j=1}^{J} jF(j)\cos(\omega_j n\Delta t + \psi_j), \tag{A27}$$

$$F(j) = j^{(\eta/2)-1}. \tag{A28}$$

As described in Ref. [46], α is an arbitrary scaling factor, ω_0 is the fundamental spacing between true adjacent discrete frequencies, such that $\omega_j = 2\pi f_0 j = \omega_0 j$, $j = 1, 2, \ldots, J$. For each frequency component, there exists a uniformly distributed random phase, $\psi_j \in [0, \pi]$. The free parameter η allows one to specify an arbitrary shape of the true power spectral density of f. In particular, the free parameters α, J, ω_0, and η are true dephasing noise parameters which any prediction algorithm cannot know beforehand.

It is straightforward to show that f is covariance stationary. To show mean-square ergodicity of f, one requires phases that are randomly uniformly distributed over one cycle for each harmonic component of f [54]. Subsequently, one shows that an ensemble average and a longtime average of a multicomponent engineered f are equal. For the evaluation of the longtime average, we use product-to-sum formulas and observe that the case $j \neq j'$ has a zero contribution, as any finite contributions from cosine terms over a symmetric integral are reduced to zero as $N \to \infty$. For $j = j'$, only a single cosine term survives. The surviving term depends on ν and N cancels to yield a finite, nonzero contribution that matches the ensemble average.

We briefly comment that f is Gaussian by the central limit theorem in the regimes considered in this paper. The probability density function of a sum of random variables is a convolution of the individual probability density functions. The central limit theorem grants that each element of f_n at n appears Gaussian distributed for large values of J, irrespective of the underlying properties of the constituent terms or the distribution of the phases ψ. A numerical analysis shows that $J > 15$ results in each f_n appearing to be approximately Gaussian distributed.

There is an important difference between f_n defined in this appendix and f_n defined in Appendixes B and C. In Appendixes B and C, the term f_n defines the "true model" for an algorithmic representation of an arbitrary covariance-stationary process—by invoking either Wold's decomposition theorem (the AKF and the QKF) or the spectral representation theorem (the LKFFB and GPR with the periodic kernel). This means that f_n in the subsequent appendixes approximates the true covariance-stationary stochastic qubit phases, $\{f_n\}$, of this appendix in the limit only where the total size of the available sample data increases to infinity. Our notation f_n fails to distinguish between these two different interpretations, as such a difference does not arise in typical applications—in our case, we have no *a priori* true model of describing stochastic qubit phases, and we must rely on mean-square approximations. Henceforth, we retain f_n to be the true model for an algorithm with the understanding that this notation refers to an approximate representation of an arbitrary, covariance-stationary sequence of stochastic qubit phases. We reserve the use of $\hat{f}_n$ for the state estimates and predictions that an algorithm makes having considered a single noisy measurement record.

APPENDIX B: AUTOREGRESSIVE REPRESENTATION OF f IN AN AKF (AND A QKF)

Our objective in this appendix is to justify the representation of f_n assumed by the AKF. In particular, we justify that any f_n drawn from any arbitrary power spectral density satisfying the properties in Appendix A 2 can be approximated by a high-order autoregressive process.

Such results are well known, if dispersed among standard engineering and econometrics textbooks [4,11,33–35,55]. We have struggled to find standard references that explicitly link high-q AR models in approximating arbitrary covariance-stationary time series of arbitrary power spectral densities, though some general comments are made in Ref. [55]. In the discussion below, we summarize relevant background material and link a high-q AR process to a theorem that guarantees arbitrary representation of zero-mean covariance-stationary processes, and we provide explicit references for proofs that are beyond the scope of the introductory remarks in this appendix. We consider AR processes of order q [AR(q)], and moving-average processes of order p [MA(p)]. A model incorporating both types of processes is known as an ARMA(q, p) model in our notation.

First, we define the lag operator, $\mathcal{L}$. This operator defines a map between time-series sequences and enables a compact description of ARMA processes. For an infinite time series $\{f_n\}_{n=-\infty}^{\infty}$ and a constant scalar, c, the lag operator is defined by the following properties:

$$\mathcal{L} f_n = f_{n-1}, \tag{B1}$$

$$\mathcal{L}^q f_n = f_{n-q}, \tag{B2}$$

$$\mathcal{L}(c f_n) = c \mathcal{L} f_n = c f_{n-1}, \tag{B3}$$

$$\mathcal{L} f_n = c, \quad \forall n, \Rightarrow \mathcal{L}^q f_n = c. \tag{B4}$$

Next, we define a Gaussian white-noise sequence, ξ, under a stronger condition than what is stated simply in Eq. (B6), that ξ_{n_1} and ξ_{n_2} are independent for all values of n_1 and n_2:

$$\mathbb{E}[\xi] \equiv 0, \tag{B5}$$

$$\mathbb{E}[\xi_{n_1} \xi_{n_2}] \equiv \sigma^2 \delta(n_1 - n_2). \tag{B6}$$

With these definitions, we can define an autoregressive process and a moving-average process of unity order. Equation (B7) defines an AR$(q = 1)$ process, and the dynamics of f are given as lagged values of f_n. The second definition in Eq. (B8) depicts a MA$(p = 1)$ process where the dynamics are given by lagged values of the Gaussian white noise ξ:

$$(1 - \phi_1 \mathcal{L}) f_n = c + \xi_n, \tag{B7}$$

$$f_n = c' + (\Psi_1 \mathcal{L} + 1)\xi_n. \tag{B8}$$

Here, Ψ_1 and ϕ_1 are known scalars defining the dynamics of f_n, w_n is a white-noise Gaussian process, and c and c' are fixed scalars. It is well known that a MA(∞) representation is equivalently an AR(1) process, and the reverse relationship also applies. For example, we can rewrite Eq. (B7) as

$$f_n = c + \xi_n + \phi_1 f_{n-1} \tag{B9}$$

$$= w_n + \phi_1 f_{n-1} \tag{B10}$$

$$= w_n + \phi_1 (w_{n-1} + \phi_1 f_{n-2}) \tag{B11}$$

$$\vdots \tag{B12}$$

$$= \phi_1^{n+1} F_0 + \phi_1^n w_0 + \phi_1^{n-1} w_1 + \cdots w_n \tag{B13}$$

$$= \phi_1^{n+1} F_0 + \phi_1^n (c + \xi_0) + \cdots + (c + \xi_n) \tag{B14}$$

$$= \phi_1^{n+1} F_0 + c(\phi_1^n + \phi_1^{n-1} + \cdots + 1) + \sum_{k=0}^{n} \phi_1^k \xi_{n-k}, \tag{B15}$$

$$w_n \equiv c + \xi_n, \tag{B16}$$

$$F_0 \equiv f_{n=-1}. \tag{B17}$$

In the last line (and for all subsequent analysis in this appendix), k should only be interpreted as a index variable for compactly rewriting terms in an equation as summations. We restrict $|\phi_1|$ to be less than 1, such that f is covariance stationary [34]. Under these conditions, we take the limit of f capturing an infinite past, namely, as $n \to \infty$. The initial state F_0 is eventually forgotten: $\phi_1^{n+1} F_0 \approx 0$ if n is large and $|\phi_1| < 1$. Similarly, the terms $c(\phi_1^n + \phi_1^{n-1} + \cdots + 1)$ can be summarized as a geometric series in ϕ_1. The remaining terms satisfy the definition of a MA(∞) process:

$$f_n = c \frac{1}{1 - |\phi_1|} + \sum_{k=0}^{\infty} \phi_1^k \xi_{n-k}, \qquad |\phi_1| < 1 \tag{B18}$$

It is straightforward to show that the reverse is true, namely, a MA(1) is equivalent to an AR(∞) representation [34].

The consideration of a MA(∞) process leads us directly to Wold's decomposition for arbitrary covariance-stationary processes, namely, that any covariance-stationary f can be represented as

$$f_n \equiv c' + \sum_{k=0}^{\infty} \Psi_k \mathcal{L}^k \xi_n, \tag{B19}$$

$$c' \equiv \mathbb{E}[f_n | f_{n-1}, f_{n-2}, \cdots], \tag{B20}$$

$$\Psi_0 \equiv 1, \tag{B21}$$

$$\sum_{k=0}^{\infty} \Psi_k^2 < \infty. \tag{B22}$$

Equation (B19) defines the MA(∞) process derived previously as an AR(1) process. This process is ergodic for a Gaussian ξ. However, such a representation of f requires fitting data to an infinite number of parameters $\{\Psi_1, \Psi_2, \cdots\}$, and approximations must be made.

We approximate an arbitrary covariance stationary f using finite but high-order AR(q) processes. Below, we show that any finite-order AR(q) process has a MA(∞) representation satisfying Wold's theorem.

We define an arbitrary AR(q) process as

$$\xi_n \equiv (1 - \phi_1 \mathcal{L} - \phi_2 \mathcal{L}^2 - \cdots - \phi_q \mathcal{L}^q)(f_n - c). \tag{B23}$$

In particular, we define λ_i, $i = 1, \ldots, q$ as eiqenvalues of the dynamical model, Φ:

$$\Phi \equiv \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{q-1} & \phi_q \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \tag{B24}$$

$$\lambda \equiv [\lambda_1 \cdots \lambda_q], \qquad \text{such that } |\Phi - \lambda \mathcal{I}_q| = 0. \tag{B25}$$

We use the following result from Ref. [34] without proof that the above implies

$$1 - \phi_1 \mathcal{L} - \phi_2 \mathcal{L}^2 - \cdots - \phi_q \mathcal{L}^q \tag{B26}$$

$$\equiv (1 - \lambda_1 \mathcal{L}), \ldots, (1 - \lambda_q \mathcal{L}). \tag{B27}$$

The equation above allows us factorize as

$$\xi_n = (1 - \lambda_1 \mathcal{L}), \ldots, (1 - \lambda_q \mathcal{L})(f_n - c). \tag{B28}$$

For us to invert this problem and recover a MA process, we need to show that the inverse for each $(1 - \lambda_{q'} \mathcal{L})$ term exists for $q' = 1, \ldots, q$. To do so, we start by defining the operator $\Lambda_q(\mathcal{L})$:

$$\Lambda_q(\mathcal{L}) \equiv \lim_{k\to\infty}(1 + \lambda_q\mathcal{L} + \cdots + \lambda_q^k\mathcal{L}^k). \tag{B29}$$

We consider an arbitrary q'th eigenvalue term in process and we multiply it by $\Lambda_{q'}(\mathcal{L})$:

$$\Lambda_{q'}(\mathcal{L})\xi_n = \Lambda_{q'}(\mathcal{L})(1-\lambda_0\mathcal{L}),\ldots,(1-\lambda_{q'}\mathcal{L}),\ldots,(f_n-c) \tag{B30}$$

$$= \lim_{k\to\infty}(1 + \lambda_{q'}\mathcal{L} + \cdots + \lambda_{q'}^k\mathcal{L}^k)(1-\lambda_{q'}\mathcal{L})(1-\lambda_0\mathcal{L}),\ldots,(1-\lambda_{q'-1}\mathcal{L})(1-\lambda_{q'+1}\mathcal{L}),\ldots,(1-\lambda_q\mathcal{L})(f_n-c) \tag{B31}$$

$$= \lim_{k\to\infty}(1 + \lambda_{q'}\mathcal{L} + \cdots + \lambda_{q'}^k\mathcal{L}^k)(1-\lambda_0\mathcal{L}),\ldots,(1-\lambda_{q'-1}\mathcal{L})(1-\lambda_{q'+1}\mathcal{L}),\ldots,(1-\lambda_q\mathcal{L})(f_n-c) \tag{B32}$$

$$- \lim_{k\to\infty}(\lambda_{q'}\mathcal{L} + \cdots + \lambda_{q'}^{k+1}\mathcal{L}^{k+1})(1-\lambda_0\mathcal{L}),\ldots,(1-\lambda_{q'-1}\mathcal{L})(1-\lambda_{q'+1}\mathcal{L}),\ldots,(1-\lambda_q\mathcal{L})(f_n-c) \tag{B33}$$

$$= \lim_{k\to\infty}(1 + \lambda_{q'}^{k+1}\mathcal{L}^{k+1})(1-\lambda_0\mathcal{L}),\ldots,(1-\lambda_{q'-1}\mathcal{L})(1-\lambda_{q'+1}\mathcal{L}),\ldots,(1-\lambda_q\mathcal{L})(f_n-c). \tag{B34}$$

Each of the residual terms $\lambda_{q'}^{k+1}\mathcal{L}^{k+1} \to 0$ if $|\lambda_{q'}| < 1$ for large values of k, and this case $\Lambda_{q'}(\mathcal{L})$ defines the inverse $(1-\lambda_{q'}\mathcal{L})^{-1}$ value. This procedure is repeated for all q eigenvalues to invert Eq. (B28) and, subsequently, perform a partial fraction expansion as follows:

$$f_n - c = \frac{1}{(1-\lambda_1\mathcal{L})\ldots(1-\lambda_q\mathcal{L})}\xi_n \tag{B35}$$

$$= \sum_{q'=1}^{q}\frac{a_{q'}}{1-\lambda_{q'}\mathcal{L}}\xi_n, \tag{B36}$$

$$a_{q'} \equiv \frac{\lambda_{q'}^{q-1}}{\prod_{q''=1,q''\neq q'}^{q}(\lambda_{q'}-\lambda_{q''})}. \tag{B37}$$

The coefficients are $a_{q'}$, as obtained via the partial fraction expansion method, during which $\mathcal{L}$ is treated as an ordinary polynomial. At present, we have to represent f via a finite-q weighted average of values of ξ. However, in substituting the definition of $\Lambda_{q'} \equiv (1-\lambda_{q'}\mathcal{L})^{-1}$ from Eq. (B29) into Eq. (B36) and regrouping the terms in powers of $\mathcal{L}$, we recover the form of a MA representation (setting $c \equiv \tilde{f}_n = 0$, $\forall\, n$, for simplicity):

$$f_n = \left[\sum_{q'=1}^{q} a_{q'}\mathcal{L}^0 + \lim_{k\to\infty}\sum_{k'=1}^{k}\left(\sum_{q'=1}^{q} a_{q'}\lambda_{q'}^{k'}\right)\mathcal{L}^{k'}\right]\xi_n \tag{B38}$$

$$= \Psi_0 + \sum_{k=1}^{\infty}\Psi_k\mathcal{L}^k\xi_n, \tag{B39}$$

$$\Psi_0 \equiv \sum_{q'=1}^{q} a_{q'}\mathcal{L}^0, \tag{B40}$$

$$\Psi_k \equiv \sum_{q'=1}^{q} a_{q'}\lambda_{q'}^{k'}. \tag{B41}$$

By examining the properties of Φ raised to arbitrary powers, it can be shown that $\sum_{q'=1}^{q} a_{q'} \equiv 1$ and that Ψ_k is the first element of Φ raised to the kth power [34], yielding the absolute summability of Ψ_k if $|\phi_{q'<q}| < 1$. These results ensure that Wold's theorem is fully satisfied and that an AR(p) process has a MA(∞) representation. In moving to an arbitrarily high value of q, we enable the approximation of any covariance stationary f.

For proofs that high-q AR approximations for covariance stationary f improve with q, see, for example, Ref. [37]. The key correspondence is that the number of finite lag terms q in an AR(q) model contribute to the first q values of the covariance function. This approximation improves with q even if f is not a true AR process [37,55]. Asymptotically efficient coefficient estimates for any MA(∞) representation of f are obtained by letting the order of a purely AR(q) process tend to infinity and increasing the total data size, N [37].

When data is fixed at N, we expect a high-q model to gradually saturate in a predictive estimation performance. One can arbitrarily increase performance by increasing both q and N [37]. In our application with finite data N, we increase q to settle on a high-order AR model while training the LSF to track arbitrary covariance-stationary power spectral densities [35].

A high-q AR model is often the first step for developing models with smaller number of parameters, for example, considering a mixture of finite-order AR(q) and MA(p) models and estimating a $p+q$ number of coefficients using a range of standard protocols [35,55]. The design of potential ARMA models for our application requires further investigation that is beyond the scope of this paper.

APPENDIX C: SPECTRAL REPRESENTATION OF f IN GPR (PERIODIC KERNEL) AND A LKFFB

The well-known spectral representation theorem guarantees that any covariance-stationary random process (real or complex) can be represented in a generalized harmonic basis. We defer a detailed treatment of spectral analysis of covariance-stationary processes to standard textbooks—for example, Refs. [34,38]—and present background and key results to provide insights into the choice of a LKFFB and GPR (periodic kernel).

The spectral representation theorem states that any covariance-stationary random process has a representation given by f_n and, correspondingly, a probability distribution, $F(\omega)$, over $[-\pi, \pi]$ in the dual domain, such that

$$f_n = \mu_f + \int_0^\pi [a(\omega)\cos(\omega n) + b(\omega)\sin(\omega n)]d\omega, \tag{C1}$$

$$R(\nu) = \int_{-\pi}^{\pi} e^{-i\omega\nu} dF(\omega). \tag{C2}$$

Here, μ_f is the true mean of the process f. The processes $a(\omega)$ and $b(\omega)$ are zero mean and serially and mutually uncorrelated; namely, $\int_{\omega_1}^{\omega_2} a(\omega)d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} a(\omega)d\omega$ and $\int_{\omega_j}^{\omega_{j'}} b(\omega)d\omega$ for any case where $\omega_1 < \omega_2 < \omega_3 < \omega_4$ and any choice of j and j' within the half cycle $[0, \pi]$.

The distribution $F(\omega)$ exists as a limiting case considering cumulative probability density functions for f_n at each n and letting $n \to \infty$, such that a sequence of these density functions approaches $F(\omega)$ [38]. If $F(\omega)$ is differentiable with respect to ω, then we see the power spectral density $S(\omega)$, and $R(\nu)$ represents the Fourier duals [38]:

$$R(\nu) = \int_{-\pi}^{\pi} e^{-i\omega\nu} S(\omega)d\omega, \tag{C3}$$

$$S(\omega) \equiv \frac{dF(\omega)}{d\omega}. \tag{C4}$$

The duality of the covariance function and the spectral density is formally expressed in the literature by the Wiener-Khinchin theorem.

We consider the finite sample analog of the spectral representation theorem considered above by following Ref. [34]. To proceed, we define mean-square convergence as a distance metric for determining when a sequence of random variables $\{\hat{f}_n\}$ converges to a random variable, f_n, in the mean-square limit if

$$\mathbb{E}[\hat{f}_n^2] < \infty \quad \forall n, \tag{C5}$$

$$\lim_{n\to\infty} \mathbb{E}[\hat{f}_n - f_n] = \lim_{n\to\infty} \|\hat{f}_n - f_n\| = 0. \tag{C6}$$

The statement $\|\hat{f}_n - f_n\| = 0$ measures the closeness between random variables $\hat{f}_n$ and f_n, even though the mean-square limit is defined for terms of a sequence of random variables, $\{\hat{f}_n\}$, where convergence improves with $n \to \infty$. In context of this work, we define $\hat{f}_n$ as a linear predictor of f_n belonging to a covariance-stationary f. Hence, each $\hat{f}_n$ for a large value of n is a linear combination of the set of random variables belonging all past noisy observations (and, in Kalman filtering, all past state predictions). Mean-square convergence of $\|\hat{f}_n - f_n\| = 0$ in our context is a statement of the quality of a predictor, $\hat{f}_n$, in predicting f_n as the total measurement data grow.

Next, we account for finite data and define the finite sample analog for the spectral representation theorem. We suppose there exists a set of arbitrary, fixed frequencies $\{\omega_j\}$ for $j = 1, \ldots, J$. We let n denote finite time steps for observing f_n at $n = 1, \ldots, N$. Furthermore, we define a set of zero-mean, mutually and serially uncorrelated random process $\{a_j\}$ and $\{b_j\}$ as finite sample analog of the true $a(\omega)$ and $b(\omega)$ values for the jth spectral component. In particular, these processes are constant over n by the covariance stationarity of f. Then, the finite sample analog for the spectral representation theorem becomes [34]

$$f_n = \mu_f + \sum_{j=1}^{J} [a_j \cos(\omega_j n) + b_j \sin(\omega_j n)], \tag{C7}$$

$$\mathbb{E}[a_j] = \mathbb{E}[b_j] = 0, \tag{C8}$$

$$\mathbb{E}[a_j a_{j'}] = \mathbb{E}[b_j b_{j'}] = \sigma_j^2 \delta(j - j'), \tag{C9}$$

$$\mathbb{E}[a_j b_{j'}] = 0 \quad \forall j, j', \tag{C10}$$

$$\mu_f \equiv 0. \tag{C11}$$

The last line enforces a zero-mean stochastic process and simplifies the analysis without loss of generality, and $\delta(\cdots)$ is the Kronecker-δ function with arguments, j, j'.

To illustrate, the first two moments are of the form

$$\mathbb{E}[f_n] = 0, \tag{C12}$$

$$R(\nu) = \sigma^2 \sum_j^J p_j \cos(\omega_j \nu), \tag{C13}$$

$$p_j \equiv \frac{\sigma_j^2}{\sigma^2} \equiv \frac{\sigma_j^2}{\sum_j \sigma_j^2}. \tag{C14}$$

We introduce measurement noise into the formula for true values of f_n, and doing so establishes a commonality with Kalman filtering for a covariance-stationary process.

An ordinary least-squares (OLS) regression can be constructed by providing a collection of $J^{(B)}$ basis frequencies $\{\omega_j^{(B)}\}$, as in Ref. [34]. The OLS problem is constructed by separating the set of coefficients $\{\hat{\mu}_f, \hat{a}_1, \hat{b}_1, \ldots, \hat{a}_J, \hat{b}_J\}$ and regressors $\{1, \cos[\omega_1(n-1)], \sin[\omega_1(n-1)], \ldots, \cos[\omega_J^{(B)}(n-1)], \sin[\omega_J^{(B)}(n-1)]\}$. For the specific choice of basis, $J^{(B)} = (N-1)/2$ (odd values of N) and $\omega_j^{(B)} \equiv 2\pi j/N$, we state the key result from Ref. [34], that the coefficient estimates are obtained as

$$\hat{f}_n = \hat{\mu}_f + \sum_{j=1}^{J^{(B)}} \{\hat{a}_j \cos[\omega_j^{(B)}(n-1)] + \hat{b}_j \sin[\omega_j^{(B)}(n-1)]\}, \tag{C15}$$

$$\hat{a}_j \equiv \frac{2}{N} \sum_{n'=1}^{N} \hat{f}_{n'} \cos[\omega_j^{(B)}(n'-1)], \tag{C16}$$

$$\hat{b}_j \equiv \frac{2}{N} \sum_{n'=1}^{N} \hat{f}_{n'} \sin[\omega_j^{(B)}(n'-1)]. \tag{C17}$$

This choice of basis results in the number of regressors being the same as the length of the measurement record. Furthermore, the term $(\hat{a}_j^2 + \hat{b}_j^2)$ is proportional to the total contribution of the jth spectral component to the total sample variance of f, or, in other words, the amplitude estimate for the power spectral density of true values of f.

Next, we depart from the OLS problem above in several ways: first, by introducing process noise and, second, by changing the basis oscillators considered in the problem above. As in the main text, the linear measurement record is defined as

$$y_n \equiv f_n + v_n, \tag{C18}$$

$$v_n \sim \mathcal{N}(0, R). \tag{C19}$$

The link in GPR (periodic kernel) is direct and the link with the LKFFB is made by setting $f_n \equiv H_n x_n$. In both frameworks, we incorporate the effect of measurement noise through the measurement-noise variance, R, which has the effect of regularizing the least-squares estimation process discussed above.

1. Infinite basis of oscillators in a GPR periodic kernel

In GPR (periodic kernel) data are projected on an infinite basis of oscillators, namely, $J^{(B)} \rightarrow \infty$.

To see this, we follow the sketch of a proof provided in Ref. [42] to show that a sine-squared exponential (periodic kernel) used in Gaussian process regression satisfies the covariance function of trigonometric polynomials. Here, the index j labels an infinite comb of oscillators, and m represents the higher-order terms in the power reduction formulas in the last line of the definition below:

$$\omega_0^{(B)} \equiv \frac{\omega_j^{(B)}}{j}, \qquad j \in \{0, 1, \ldots, J^{(B)}\}, \tag{C20}$$

$$R(\nu) \equiv \sigma^2 \exp\left(-\frac{2\sin^2(\frac{\omega_0^{(B)}\nu}{2})}{l^2}\right) \tag{C21}$$

$$= \sigma^2 \exp\left(-\frac{1}{l^2}\right) \exp\left(\frac{\cos(\omega_0^{(B)}\nu)}{l^2}\right) \tag{C22}$$

$$= \sigma^2 \exp\left(-\frac{1}{l^2}\right) \sum_{m=0}^{M\rightarrow\infty} \frac{1}{m!} \frac{\cos^m(\omega_0^{(B)}\nu)}{l^{2m}}. \tag{C23}$$

Next, we expand each cosine using power reduction formulas for odd and even powers, respectively, and we regroup the terms. For example, we expand the terms for $m = 0, 1, 2, 3, 4, 5, \ldots$ as

$$R(\nu) = \sigma^2 \exp\left(-\frac{1}{l^2}\right)\cos(\omega_0^{(B)}\nu)\left[\frac{2}{(2l^2)}\binom{1}{0} + \frac{2}{(2l^2)^3}\frac{1}{3!}\binom{3}{1} + \frac{2}{(2l^2)^5}\frac{1}{5!}\binom{5}{2}\cdots\right] \tag{C24}$$

$$+\sigma^2 \exp\left(-\frac{1}{l^2}\right)\cos(2\omega_0^{(B)}\nu)\left[\frac{2}{(2l^2)^2}\frac{1}{2!}\binom{2}{0} + \frac{2}{(2l^2)^4}\frac{1}{4!}\binom{4}{1} + \cdots\right] \tag{C25}$$

$$+\sigma^2 \exp\left(-\frac{1}{l^2}\right)\cos(3\omega_0^{(B)}\nu)\left[\frac{2}{(2l^2)^3}\frac{1}{3!}\binom{3}{0} + \frac{2}{(2l^2)^5}\frac{1}{5!}\binom{5}{1}\cdots\right] \tag{C26}$$

$$+\sigma^2 \exp\left(-\frac{1}{l^2}\right)\cos(4\omega_0^{(B)}\nu)\left[\frac{2}{(2l^2)^4}\frac{1}{4!}\binom{4}{0} + \cdots\right] \tag{C27}$$

$$+\sigma^2 \exp\left(-\frac{1}{l^2}\right)\cos(5\omega_0^{(B)}\nu)\left[\frac{2}{(2l^2)^5}\frac{1}{5!}\binom{5}{0} + \cdots\right] \tag{C28}$$

$$\vdots$$

$$+\sigma^2 \exp\left(-\frac{1}{l^2}\right)\left[\frac{1}{(2l^2)^2}\frac{1}{2!}\binom{2}{1} + \frac{1}{(2l)^4}\frac{1}{4!}\binom{4}{2} + \cdots\right] + \sigma^2 \exp\left(-\frac{1}{l^2}\right). \tag{C29}$$

In the expansion above, the vertical and horizontal dots represent contributions from $m > 5$ terms. The key message is that truncating m to a finite number of terms M truncates j to represent a finite number of oscillators. For the example above, if the power reduction expansion indexed by m above is truncated to $M = 5$ terms, then the number of basis oscillators (the number of rows) is also be truncated. We now summarize the amplitude equations (C24)–(C28) in the second term of $R(\nu)$, and Eq. (C29) corresponds to the $p_{0,M}$ term below:

$$R(\nu) = \sigma^2\left(p_{0,M} + \sum_{j=0}^{\infty} p_{j,M}\cos(j\omega_0^{(B)}\nu)\right), \tag{C30}$$

$$p_{j,M} \equiv \sigma^2 \exp\left(-\frac{1}{l^2}\right)\sum_{\beta=0}^{\beta=\beta_{j,m}^{\max}} \frac{2}{(2l^2)^{(j+2\beta)}}\frac{1}{(j+2\beta)!}\binom{j+2\beta}{\beta}, \tag{C31}$$

$$\beta \equiv 0, 1, \ldots, \beta_{j,m}^{\max}, \tag{C32}$$

$$p_{0,M} = \exp\left(-\frac{1}{l^2}\right)\sum_{\alpha=0}^{\alpha=\alpha_m^{\max}} \frac{1}{(2l^2)^{(2\alpha)}}\frac{1}{(2\alpha)!}\binom{2\alpha}{\alpha}, \tag{C33}$$

$$\alpha \equiv 0, 1, \ldots, \alpha_m^{\max}. \tag{C34}$$

By examining the cosine expansion, one sees that a truncation at $(M, J^{(B)})$ means our summarized formulas require $\beta_{j,M}^{\max} = \lfloor (M - j)/2 \rfloor$ and $\alpha_M^{\max} = \lfloor (M/2) \rfloor$ where $\lfloor \cdots \rfloor$ denotes the ceiling floor. If we truncate using $M \equiv J^{(B)}$ such that $\alpha_M^{\max} = \lfloor (J^{(B)}/2) \rfloor$ and $\beta_{j,M}^{\max} = \lfloor (J^{(B)} - j)/2 \rfloor$ and readjust the kernel for the zeroth frequency term, then we agree with the final result in Ref. [42].

We compare the covariance function of the periodic kernel in Eq. (C30) with the covariance function of the trigonometric polynomials in Eq. (C13). Here, the $p_{j,M}$ values for the periodic kernel are not identically specified, in general, to those under the spectral representation theorem, but they otherwise retain a structure as a cosine basis where the correlations between two random variables in a sequence depends only on the separation between them. For a constant-mean Gaussian process, the form of the periodic kernel allows the underlying process to satisfy covariance stationarity and appears to permit an interpretation via the spectral representation theorem.

2. Amplitude and phase extraction for the finite oscillator basis in the LKFFB

In the LKFFB, we specify a fixed basis of oscillators at the physical Fourier resolution established by the measurement record. Using a specific state-space model, we can track amplitudes and phases for each basis oscillator individually to enable forward prediction at any time step of our choosing. The design of a fixed basis necessarily incorporates prior assumptions about the extent to which a fast measurement action oversamples slowly drifting non-Markovian noise, that is, a (potentially incorrect) assumption about dephasing noise bandwidth.

The efficacy of the LKFFB in our application assumes an appropriate choice of the "Kalman basis" oscillators. The choice of basis can effect the forward prediction of the state estimates. To illustrate, consider the choice of bases A–C defined below. Basis A depicts a constant spacing above the Fourier resolution (e.g., $\omega_0^{(B)} \geq [(2\pi)/(N_T\Delta t)]$). Basis B introduces a minimum Fourier resolution and effectively creates an irregular spacing if one wishes to consider a basis frequency comb coarser than the experimentally established Fourier spacing over the course of the experiment. Basis C is identical to basis B but allows a projection to a zero frequency component,

$$\text{basis } A\colon \equiv \{0, \omega_0^{(B)}, 2\omega_0^{(B)}, \ldots, J^{(B)}\omega_0^{(B)}\}, \tag{C35}$$

$$\text{basis } B\colon \equiv \left\{\frac{2\pi}{N\Delta t}, \frac{2\pi}{N\Delta t} + \omega_0^{(B)}, \ldots, \frac{2\pi}{N\Delta t} + J^{(B)}\omega_0^{(B)}\right\}, \tag{C36}$$

$$\text{basis } C\colon \equiv \left\{0, \frac{2\pi}{N\Delta t}, \frac{2\pi}{N\Delta t} + \omega_0^{(B)}, \ldots, \frac{2\pi}{N\Delta t} + J^{(B)}\omega_0^{(B)}\right\}. \tag{C37}$$

While one can propagate the LKFFB with zero gain, it may be advantageous for predictive control applications to generate predictions in one calculation, rather than recursively. This means that we sum contributions over all $j \in J^{(B)}$ oscillators, and we reconstruct the signal for all future time values in one calculation, without having to propagate the filter recursively with zero gain. The interpretation of the predicted signal, $\hat{f}_n$, requires an additional (but time-constant) phase correction term ψ_C that arises as a byproduct of the computational basis (i.e., basis A, B, or C). The phase correction term corrects for a gradual misalignment between the Fourier and computational grids which occurs if one specifies a nonregular spacing inherent in basis B or C. Let n_C denote the time step at which instantaneous amplitudes $\|\hat{x}^j_{n_C}\|$ and instantaneous phase $\theta_{\hat{x}^j_{n_C}}$ is extracted for the oscillator represented by the jth state-space resonator, x^j_n, where the superscript j denotes an oscillator of frequency $\omega_j^{(B)} \equiv j\omega_0^{(B)}$ (not a power):

$$\hat{f} = \sum_{j=0}^{J^{(B)}} \|\hat{x}^j_{n_C}\| \cos(m\Delta t\omega_j^{(B)} + \theta_{\hat{x}^j_{n_C}} + \psi_C), \tag{C38}$$

$$n_C \in N_T, \qquad m \in N_P,$$

$$\psi_C \equiv \begin{cases} 0 & \text{(basis } A) \\ \frac{2\pi}{\omega_0^{(B)}}\left(\omega_0^{(B)} - \frac{2\pi}{N\Delta t}\right) & \text{(basis } B \text{ or } C) \end{cases}. \tag{C39}$$

The output predictions from calculating a harmonic sum using learned instantaneous amplitudes and phases and the LKFFB bases A–C agree with zero-gain predictions if ψ_C is specified as above. The calculation of ψ_C is determined entirely by the choice of computational and experimental sampling procedures, and it assumes no information about true dephasing.

Next, we define an analytical ratio to define the optimal training time, n_C, at which LKFFB predictions should commence, irrespective of whether the prediction procedure recursively propagates the Kalman filter with zero gain, or by calculating a harmonic sum for all prediction points in one go:

$$n_C \equiv \frac{2\pi}{\Delta t\omega_0^{(B)}} = \frac{2\pi f_s}{\omega_0^{(B)}}. \tag{C40}$$

Consider an arbitrarily chosen training period, $N_T \neq n_C$. For fixed values of f_s, our choice of $N_T > n_C$ means we are achieving a Fourier resolution which exceeds the resolution of the LKFFB basis. Now consider that $N_T < n_C$. This means that we have extracted information prematurely, and we have not waited long enough to project on the smallest basis frequency, namely, $\omega_0^{(B)}$. In the case where the data are perfectly projected on our basis, the choice of n_C has no impact. For imperfect learning, we see that instantaneous amplitude and phase information slowly degrades for $N_T > n_C$, and trajectories for the smallest basis frequency are not stabilized for $N_T < n_C$.

Of these choices, basis A for $\omega_0^{(B)} \equiv [(2\pi)/(N_T\Delta t)]$ is expected to yield the best performance, at the expense of computational load, as confirmed in numerical experiments. All results in this paper are reported for basis A with $N_T \equiv n_C$.

3. Equivalent spectral representation of f in the LKFFB and the GPR periodic kernel

In this section, we consider the structural similarities between the LKFFB and GPR with a periodic kernel. We show that the LKFFB has an structure analogous to a stack of stochastic processes on a circle [38], and, in moving from discrete to continuous time, we recover a covariance function that has the same structure if the periodic kernel is truncated to a finite basis of oscillators, $J^{(B)}$. For zero-mean, Gaussian random variables, covariance stationarity is established, completing the link between the LKFFB and the periodic kernel. For the case $\Gamma_n w_n \rightarrow w_n$ in the LKFFB, stacked Kalman resonators as an approximation to infinite oscillators in a periodic kernel is documented in Ref. [42].

At time step n, the posterior Kalman state at $n-1$ acts as the initial state at n, such that $\nu = \Delta t$ for a small Δt such that a linearized trajectory is approximately true for each basis frequency. Using the following correlation relations and a Gaussian assumption for process noise, where $n, m \in N$ are indices for time steps and $j = 0, 1, \ldots, J^{(B)}$ indexes the set of basis oscillators:

$$\mathbb{E}[w_n] = 0 \quad \forall j \in J^{(B)}, \quad n \in N, \tag{C41}$$

$$\mathbb{E}[w_n, w_m] = \sigma^2 \delta(n-m), \quad n, m \in N, \tag{C42}$$

$$\mathbb{E}[A_0^j] = \mathbb{E}[B_0^{j'}] = 0 \quad \forall j, j' \in J^{(B)}, \tag{C43}$$

$$\mathbb{E}[A_n^j B_m^{j'}] = 0 \quad \forall j, j' \in J^{(B)}, \quad n, m \in N, \tag{C44}$$

$$\begin{aligned}\mathbb{E}[A_n^j A_m^{j'}] = \mathbb{E}[B_n^j B_m^{j'}] = \sigma_j^2 \delta(n-m)\delta(j-j') \\ \forall j, j' \in J^{(B)}, \quad n, m \in N,\end{aligned} \tag{C45}$$

$$\mathbb{E}[w_n A_m^j] = \mathbb{E}[w_n B_m^{j'}] \equiv 0 \quad \forall j, j' \in J^{(B)}, \quad n, m \in N. \tag{C46}$$

Consider a jth state-space resonator, x_n^j, in the LKFFB, where the superscript j denotes an oscillator (not a power), and we obtain

$$\Theta(j\omega_0^{(B)}\Delta t) = \begin{bmatrix} \cos(j\omega_0^{(B)}\Delta t) & -\sin(j\omega_0^{(B)}\Delta t) \\ \sin(j\omega_0^{(B)}\Delta t) & \cos(j\omega_0^{(B)}\Delta t) \end{bmatrix}, \tag{C47}$$

$$x_n^j \equiv \begin{bmatrix} A_n^j \\ B_n^j \end{bmatrix} = \Theta(j\omega_0^{(B)}\Delta t)\left[\hat{\mathcal{I}} + \frac{w_{n-1}}{\sqrt{{A_{n-1}^j}^2 + {B_{n-1}^j}^2}}\right]\begin{bmatrix} A_{n-1}^j \\ B_{n-1}^j \end{bmatrix}, \tag{C48}$$

$$\Rightarrow \mathbb{E}[x_n^j] = 0, \tag{C49}$$

$$\Rightarrow \mathbb{E}[x_n^j x_m^{j\,T}]_j = \sigma_j^2 \delta(n-m) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{C50}$$

The cross-correlation terms disappear under the temporal correlation functions so defined; namely, if we assume that $n \geq m$, then states A_{m-1}^j and B_{m-1}^j at $m-1$ have, at most, a w_{n-2} term (for the case $n = m$) and cannot be correlated with a future noise term w_{n-1}.

The dynamical trajectory in the LKFFB is linearized for a small Δt. The linearization is an approximation to a true, continuous-time deterministic trajectory defining a stochastic process on a circle.

We briefly examine this continuous-time trajectory to specify the link between the LKFFB and GPR (periodic kernel). Let t denote the continuous-time deterministic dynamics for a random initial state given by a_0^j and b_0^j, where the superscript j denotes an oscillator with a frequency $\omega_j \equiv j\omega_0^{(B)}$ (not a power):

$$\mathbb{E}[a_0^j] = \mathbb{E}[b_0^{j'}] = 0 \quad \forall j, j' \in J^{(B)}, \tag{C51}$$

$$\mathbb{E}[a_0^j b_0^{j'}] = 0 \quad \forall j, j' \in J^{(B)}, \tag{C52}$$

$$\mathbb{E}[a_0^j a_0^{j'}] = \mathbb{E}[b_0^j b_0^{j'}] = \sigma_j^2 \delta(j-j') \quad \forall j, j' \in J^{(B)}, \tag{C53}$$

$$x^j(t) \equiv \begin{bmatrix} \cos(\omega_j t) & -\sin(\omega_j t) \\ \sin(\omega_j t) & \cos(\omega_j t) \end{bmatrix} \begin{bmatrix} a_0^j \\ b_0^j \end{bmatrix}, \tag{C54}$$

$$\mathbf{E}[x^j(t)] = 0, \tag{C55}$$

$$\mathbf{E}[x^j(t) x^j(t')^T] = \sigma_j^2 \begin{bmatrix} \cos(\omega_j \nu) & 0 \\ 0 & \cos(\omega_j \nu) \end{bmatrix}, \quad \nu \equiv |t' - t|. \tag{C56}$$

We see that the initial state variables, a_0^j and b_0^j, must be zero-mean independent and identically distributed variables for each value of j, such that $x^j(t)$ is covariance stationary. If a_0^j and b_0^j are Gaussian, then the joint distribution, $x^j(t)$, remains Gaussian under the linear operations above. Hence, the continuous-time limit of the dynamics in the LKFFB for $J^{(B)}$ independent substates, $x^j(t)$, describes a process with the same first and second moments for a periodic kernel truncated at $J^{(B)}$. For Gaussian processes, the LKFFB for $J^{(B)}$ stacked resonators approximately matches an expansion of the periodic kernel truncated at $J^{(B)}$.

While the formalism of the LKFFB shares a common structure with GPR (periodic kernel) in a particular limit, the physical interpretation of A_n^j and B_n^j in LKFFB is that these are components of the Hilbert transform of the original signal [29]. These components give us the ability to track and extract an instantaneous amplitude and phase associated with each basis oscillator in the LKFFB. By contrast, the coefficients of the periodic kernel are always contingent on the arbitrary truncation of the infinite basis, as seen in Eqs. (C30), (C31), and (C33). Hence, tracking (or extracting) amplitudes and phases for individual oscillators does not seem appropriate for the periodic kernel, as these values change depending on the arbitrary choice of a truncation point.

[1] J. J. J. Groen, R. Paap, and F. Ravazzolo, Real-time inflation forecasting in a changing world, J. Bus. Econ. Stat. **31**, 29 (2013).

[2] Y. Dong, Y. Li, M. Xiao, and M. Lai, Unscented Kalman filter for time varying spectral analysis of earthquake ground motions, Appl. Math. Model. **33**, 398 (2009).

[3] J. Ko and D. Fox, GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models, Auton. Robots **27**, 75 (2009).

[4] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter* (Cambridge University Press, Cambridge, England, 1990).

[5] C. Cheng, A. Sa-Ngasoongsong, O. Beyca, T. Le, H. Yang, Z. Kong, and S. T. Bukkapatnam, Time series forecasting for nonlinear and nonstationary processes: A review and comparative study, IIE Trans. **47**, 1053 (2015).

[6] J. D. Garcia and G. C. Amaral, An optimal polarization tracking algorithm for lithium-niobate-based polarization controllers, in *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Rio de Janeiro, 2016* (IEEE, New York, 2016), pp. 1–5, DOI: 10.1109/SAM.2016.7569610.

[7] F. R. Bach and M. I. Jordan, Learning graphical models for stationary time series, IEEE Trans. Signal Process. **52**, 2189 (2004).

[8] S. Tatinati and K. C. Veluvolu, A hybrid approach for short-term forecasting of wind speed, Sci. World J. **2013**, 548370 (2013).

[9] J. Hall, C. E. Rasmussen, and J. Maciejowski, Reinforcement learning with reference tracking control in continuous state spaces, in *Proceedings of the 50th IEEE Decision and Control and European Control Conference (CDC-ECC), Orlando, 2011* (IEEE, New York, 2011), pp. 6019–6024, DOI: 10.1109/CDC.2011.6161108.

[10] F. Hamilton, T. Berry, and T. Sauer, Ensemble Kalman Filtering without a Model, Phys. Rev. X **6**, 011021 (2016).

[11] J. V. Candy, *Bayesian Signal Processing: Classical, Modern, and Particle Filtering Methods*, Vol. 54 (John Wiley & Sons, Hoboken, NJ, 2016).

[12] B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications, Proc. IEEE **80**, 540 (1992).

[13] L. Ji and Z. Tie, On gradient descent algorithm for generalized phase retrieval problem, in *Proceedings of the IEEE 13th International Conference on Signal Processing (ICSP), Chengdu, China* (IEEE, New York, 2016), pp. 320–325, DOI: 10.1109/ICSP.2016.7877848.

[14] G. I. Struchalin, I. A. Pogorelov, S. S. Straupe, K. S. Kravtsov, I. V. Radchenko, and S. P. Kulik, Experimental adaptive quantum tomography of two-qubit states, Phys. Rev. A **93**, 012103 (2016).

[15] A. Sergeevich, A. Chandran, J. Combes, S. D. Bartlett, and H. M. Wiseman, Characterization of a qubit Hamiltonian using adaptive measurements in a fixed basis, Phys. Rev. A **84**, 052315 (2011).

[16] D. H. Mahler, L. A. Rozema, A. Darabi, C. Ferrie, R. Blume-Kohout, and A. M. Steinberg, Adaptive Quantum State Tomography Improves Accuracy Quadratically, Phys. Rev. Lett. **111**, 183601 (2013).

[17] M. P. V. Stenberg, O. Köhn, and F. K. Wilhelm, Characterization of decohering quantum systems: Machine learning approach, Phys. Rev. A **93**, 012122 (2016).

[18] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi, and A. G. White, Efficient Measurement of Quantum Dynamics via Compressive Sensing, Phys. Rev. Lett. **106**, 100401 (2011).

[19] Z. Shen, W. X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, Reconstructing propagation networks with natural diversity and identifying hidden sources, Nat. Commun. **5**, 4323 (2014).

[20] L. E. de Clercq, R. Oswald, C. Flühmann, B. Keitch, D. Kienzler, H.-Y. Lo, M. Marinelli, D. Nadlinger, V. Negnevitsky, and J. P. Home, Estimation of a general time-dependent Hamiltonian for a single qubit, Nat. Commun. **7**, 11218 (2016).

[21] D. Tan, S. J. Weber, I. Siddiqi, K. Mølmer, and K. W. Murch, Prediction and Retrodiction for a Continuously Monitored Superconducting Qubit, Phys. Rev. Lett. **114**, 090403 (2015).

[22] Y. Huang and J. E. Moore, Neural network representation of tensor network and chiral states, arXiv:1701.06246.

[23] C. Bonato, M. S. Blok, H. T. Dinani, D. W. Berry, M. L. Markham, D. J. Twitchen, and R. Hanson, Optimized quantum sensing with a single electron spin using real-time adaptive measurements, Nat. Nanotechnol. **11**, 247 (2016).

[24] N. Wiebe, C. Granade, A. Kapoor, and K. M. Svore, Bayesian inference via rejection filtering, arXiv:1511.06458.

[25] M. D. Shulman, S. P. Harvey, J. M. Nichol, S. D. Bartlett, A. C. Doherty, V. Umansky, and A. Yacoby, Suppressing qubit dephasing using real-time Hamiltonian estimation, Nat. Commun. **5**, 5156 (2014).

[26] C. Granade, J. Combes, and D. Cory, Practical Bayesian tomography, New J. Phys. **18**, 033024 (2016).

[27] P. E. Jacob, S. M. M. Alavi, A. Mahdi, S. J. Payne, and D. A. Howey, Bayesian inference in non-Markovian state-space models with applications to battery fractional-order systems, IEEE Trans. Control Syst. Technol. **26**, 497 (2017).

[28] S. Mavadia, V. Frey, S. D. Jarrah Sastrawan, and M. J. Biercuk, Prediction and real-time compensation of qubit decoherence via machine learning, Nat. Commun. **8**, 14106 (2017).

[29] J. Liška and E. Janeček, Time-frequency representation of instantaneous frequency using a Kalman filter, in *Robotics Automation and Control*, edited by Pavla Pecherková (InTech, Vienna, 2008), pp. 28–38, https://www.intechopen.com/books/robotics-automation-and-control.

[30] C. Ferrie, C. E. Granade, and D. G. Cory, How to best sample a periodic probability distribution or on the accuracy of Hamiltonian finding strategies, Quantum Inf. Process. **12**, 611 (2013).

[31] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, 2001).

[32] S.-M. Moon, D. G. Cole, and R. L. Clark, Real-time implementation of adaptive feedback and feedforward generalized predictive control algorithm, J. Sound Vib. **294**, 82 (2006).

[33] I. D. Landau, R. Lozano, M. M'Saad, and A. Karimi, *Adaptive Control: Algorithms, Analysis and Applications*, Vol. 51 (Springer, Berlin, 1998).

[34] J. D. Hamilton, *Time Series Analysis*, Vol. 2 (Princeton University Press, Princeton, NJ, 1994).

[35] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting* (Springer-Verlag, New York, 1996).

[36] M. Salzmann, P. Teunissen, and M. Sideris, Detection and modelling of coloured noise for Kalman filter applications,

in *Kinematic Systems in Geodesy, Surveying, and Remote Sensing*, Vol. 107, edited by K.-P. Schwarz and G. Lachapelle (Springer-Verlag, New York, 1991), pp. 251–260.

[37] B. Wahlberg, Estimation of autoregressive moving-average models via high-order autoregressive approximations, J. Time Ser. Anal. **10**, 283 (1989).

[38] S. Karlin and H. Taylor, *A First Course in Stochastic Processes* (Academic Press, New York, 1975).

[39] R. Karlsson and F. Gustafsson, Filtering and estimation for quantized sensor information, Linköping University Report No. LiTH-ISY-R-2674, 2005.

[40] B. Widrow, I. Kollar, and M. C. Liu, Statistical theory of quantization, IEEE Trans. Instrum. Meas. **45**, 353 (1996).

[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, MA, 2005).

[42] A. Solin and S. Särkkä, Explicit link between periodic covariance functions and state space models, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, 2014*, edited by S. Kaski and J. Corander (PMLR, Reykjavik, 2014), pp. 904–912.

[43] F. Tobar, T. D. Bui, and R. E. Turner, Learning stationary time series using Gaussian processes with nonparametric kernels, in *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Vol. 28 (Curran Associates, Inc., New York, 2015), pp. 3501–3509.

[44] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, Gaussian processes for time-series modelling, Phil. Trans. R. Soc. A **371**, 20110550 (2013).

[45] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging* (Springer Science+Business Media, New York, 1999).

[46] A. Soare, H. Ball, D. Hayes, X. Zhen, M. C. Jarratt, J. Sastrawan, H. Uys, and M. J. Biercuk, Experimental bath engineering for quantitative studies of quantum control, Phys. Rev. A **89**, 042329 (2014).

[47] Sheffield Machine Learning Group, GPy: A Gaussian process framework in PYTHON, http://github.com/SheffieldML/GPy, 2012.

[48] S. Arlot and P. Massart Data-driven calibration of penalties for least-squares regression, J. Mach. Learn. Res. **10**, 245 (2009), http://www.jmlr.org/papers/volume10/arlot09a/arlot09a.pdf.

[49] K. Vu, J. C. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Müller, and K. Burke, Understanding kernel ridge regression: Common behaviors from simple functions to density functionals, Int. J. Quantum Chem. **115**, 1115 (2015).

[50] P. Abbeel, A. Coates, M. Montemerlo, A. Y. Ng, and S. Thrun, Discriminative training of Kalman filters, in *Robotics: Science and Systems I*, edited by S. Thrun, G. S. Sukhatme, and S. Schaal (MIT Press, Cambridge, MA, 2005), pp. 289–296.

[51] A. Robertson and C. Grenade (unpublished).

[52] A. Wilson and R. Adams, Gaussian process kernels for pattern discovery and extrapolation, in *Proceedings of the 30th International Conference on Machine Learning (ICML '13), Atlanta, 2013*, edited by S. Dasgupta and D. McAllester, Vol. 28(3) (PMLR, Atlanta, 2013), pp. 1067–1075, http://proceedings.mlr.press/v28/wilson13.pdf.

[53] J. Quiñonero Candela, C. E. Rasmussen, A. R. Figueiras-Vidal, and M. Lázaro-Gredilla, Sparse spectrum Gaussian process regression, J. Mach. Learn. Res. **11**, 1865 (2010), http://www.jmlr.org/papers/v11/lazaro-gredilla10a.html.

[54] A. Gelb, *Applied Optimal Estimation* (MIT Press, Cambridge, MA, 1974).

[55] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models* (Springer-Verlag, New York, 1996).