Fallani, Alessio; Rossi, Matteo A.C.; Tamascelli, Dario; Genoni, Marco G.

Learning Feedback Control Strategies for Quantum Metrology

Published in:
PRX Quantum

DOI:
10.1103/PRXQuantum.3.020310

Published: 14/04/2022

Document Version
Publisher's PDF, also known as Version of record

Learning Feedback Control Strategies for Quantum Metrology

Alessio Fallani,[1] Matteo A. C. Rossi,[2,3,4] Dario Tamascelli,[1] and Marco G. Genoni[1,*]

[1] *Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, 20133 Milan, Italy*
[2] *InstituteQ - the Finnish Quantum Institute, Aalto University, Finland*
[3] *QTF Centre of Excellence, Department of Applied Physics, Aalto University, FI-00076 Aalto, Finland*
[4] *Algorithmiq Ltd., Kanavakatu 3C, FI-00160 Helsinki, Finland*

(Received 3 November 2021; revised 3 February 2022; accepted 17 March 2022; published 14 April 2022)

We consider the problem of frequency estimation for a single bosonic field evolving under a squeezing Hamiltonian and continuously monitored via homodyne detection. In particular, we exploit reinforcement learning techniques to devise feedback control strategies achieving increased estimation precision. We show that the feedback control determined by the neural network greatly surpasses in the long-time limit the performances of both the "no-control" strategy and the standard "open-loop control" strategy, which we considered as benchmarks. We indeed observe how the devised strategy is able to optimize the non-trivial estimation problem by preparing a large fraction of trajectories corresponding to more sensitive quantum conditional states.

DOI: 10.1103/PRXQuantum.3.020310

I. INTRODUCTION

The goal of quantum metrology is to devise strategies able to exploit purely quantum properties, such as entanglement and squeezing, to estimate parameters with a precision beyond that obtainable via classical means [1,2]. In the classical domain it is usual to study estimation strategies based on the continuous monitoring of a system, leading to sensors that have applications ranging from engineering to medicine.

This kind of approach is particularly interesting in the context of quantum metrology with continuously monitored quantum systems [3,4]. The role of continuous measurements is indeed twofold: on the one hand, as happens classically, the measurement output is exploited to acquire information on the parameters characterizing the system; on the other hand, the act of measuring alters the state of the system itself, thus opening the possibility of dynamically preparing more-sensitive quantum probes. Several studies have been published, both discussing the fundamental statistical tools to assess the precision achievable in this framework [5–12], and presenting practical estimation strategies [13–30].

Moreover, in the context of continuously monitored quantum systems, it is also natural to study strategies able to exploit this information to steer the evolution toward a desired quantum state via feedback control [3,31]. Much effort has been devoted to the design of strategies able to generate metrologically relevant quantum states, such as squeezed states [32–41], or to cool optomechanical systems toward their ground state, with outstanding experimental results recently reported in Refs. [42–44].

Reinforcement learning (RL) is one of the main paradigms of machine learning, together with supervised and unsupervised learning. In RL an agent learns how to perform a task by acting on a system and updating its policy through a reward-punishment mechanism [45]. The introduction of deep neural networks in RL has led to formidable results: machines have, for example, learned how to play video games [46] or how to beat expert human players at complex board games such as Go [47].

RL has recently been applied in the context of quantum information, and more generally to quantum technology, to find optimal strategies for some designated tasks [48], ranging from optimizing feedback for quantum error correction [49] to quantum control strategies [50,51], and from optimizing quantum transport [52,53] to quantum compiling [54], or even to solve Rubik's cube by exploiting quantum mechanics [55]. Recently, RL has also been used to optimize feedback control protocols in continuously monitored quantum systems [56–58], with a main focus on quantum state engineering.

*marco.genoni@fisica.unimi.it

Discovering feedback strategies, where decisions are based on previously observed measurement results, is indeed a challenging task. The stochastic nature of the problem, together with the presence of feedback mechanisms, leads to a double-exponential growth of the space of possible strategies with respect to the number of time steps. Such a task falls therefore beyond the scope of standard optimal control, and also supervised learning, techniques [57]. On the other hand, it suits the RL paradigm: the agent explores the problem space by performing random experiments on the system while learning, at the same time, an action policy.

In this work we exploit RL to design a feedback strategy optimizing a given nontrivial metrological problem. In particular, we consider the estimation of the frequency of a harmonic oscillator subjected to a squeezing Hamiltonian and undergoing a continuous homodyne detection. Differently form previous work [56–58], where the goal was the preparation of a given target state to be exploited in a selected quantum information protocol, in this work we aim to optimize real-time feedback for quantum metrology purposes, so as to attain a high precision in parameter estimation without targeting the preparation of a precise quantum state.

We show that the feedback strategy determined by RL provides a high precision in parameter estimation, and overcomes the performance of some benchmark approaches. Interestingly, the feedback protocol determined by the agent optimizes the interplay between the squeezing direction and the displacement. Given the stochastic nature of the dynamics induced by the measurement back-action, such a strategy is highly nontrivial and cannot be easily obtained with standard optimal control techniques or supervised learning techniques.

This paper is organized as follows: In Sec. II we present the physical model and the estimation problem. In Sec. III we introduce the figures of merit that we will use to assess the protocols and we discuss the role of squeezing and feedback in the estimation procedure. In Sec. IV we show how we apply RL to our problem, and in Sec. V we present our main results. We conclude the paper in Sec. VI with a brief discussion and some outlooks.

II. THE ESTIMATION PROBLEM

We consider a single bosonic mode described by the quadrature operators $(\hat{q},\hat{p})$ satisfying the canonical commutation relation $[\hat{q},\hat{p}] = i\mathbb{1}$ [59]. The evolution of the mode is determined by the Hamiltonian

$$\hat{H}_0 = \omega\hat{a}^\dagger\hat{a} + \chi(\hat{a}^2 + \hat{a}^{\dagger 2}), \tag{1}$$

with $\hat{a} = (\hat{q} + i\hat{p})/\sqrt{2}$ denoting the annihilation operator. The first term simply corresponds to the usual free quantum oscillator Hamiltonian, characterized by frequency ω; the second is a single-mode squeezing term able to generate, for $\omega = 0$ and $\chi > 0$, squeezing in the $\hat{q}$ quadrature. We recall here that a quantum state is said to be squeezed if it presents fluctuations of a quadrature operator below the vacuum shot noise. The amount of squeezing of a quantum state ϱ, for example, for the $\hat{q}$ quadrature, is typically evaluated in decibels (dB) according to the formula

$$\xi_{\rm dB} = -10\log_{10}(\langle\Delta\hat{q}^2\rangle/\langle\Delta\hat{q}^2\rangle_0), \tag{2}$$

where we have denoted with $\langle\Delta\hat{q}^2\rangle$ and $\langle\Delta\hat{q}^2\rangle_0 = 1/2$ the variance of $\hat{q}$ evaluated, respectively, for the quantum state ϱ and for the vacuum state $|0\rangle$. More generally, the maximum amount of squeezing of a single-mode quantum state along a generic quadrature operator can be evaluated as $\xi_{\rm dB} = -10\log_{10}\lambda_-$, where λ_- denotes the minimum eigenvalue of its covariance matrix $\boldsymbol{\sigma}$ (see Appendix B for more details on covariance matrices and the Gaussian formalism).

Physically, the Hamiltonian (1) describes an optical parametric oscillator that is a cavity mode with resonance frequency ω_c interacting with a nonlinear crystal and driven by a laser with frequency ω_l. It can be obtained by going to a frame rotating at the laser frequency, with $\omega = \omega_c - \omega_l$ denoting the detuning between the cavity resonance and the laser. In what follows we focus on the problem of estimating the fixed, but unknown, value of this detuning parameter ω. This kind of estimation problem was recently discussed in the standard open-system scenario for a circuit-QED implementation by also considering the usefulness of an extra Kerr-type nonlinearity in Ref. [60]. In the continuous-variable scenario one typically considers the estimation of an optical phase accumulated during a finite time evolution [61,62]. Phase estimation and frequency estimation are, however, fundamentally equivalent, and we focus on the latter as in our setup we have to deal with with a time-continuous evolution. While we phrase our results in terms of a quantum optical scenario, we expect that our findings can be extended to other physical platforms where frequency estimation is at the basis of quantum enhanced atomic clocks [63,64] and quantum magnetometry [65]. In our setting, the cavity mode is subjected to loss at rate κ. The output (leaking field) signal is then measured by means of a continuous homodyne measurement, performed with efficiency η (this parameter $\eta = \eta_D\eta_L$ takes into account both the homodyne detector efficiency η_D and the fraction of the output field η_L that is not collected by the detector). The corresponding continuous measurement outcome can be written as

$$dy_t = \sqrt{\eta\kappa}\langle\hat{a} + \hat{a}^\dagger\rangle_c\, dt + dw_t, \tag{3}$$

where $\langle\cdot\rangle_c = {\rm Tr}[\varrho_c\cdot]$ denotes the expectation over the conditional state ϱ_c, and dw_t is a Wiener increment, characterized by $\mathbb{E}[dw_t] = 0$ and $\mathbb{E}[dw_t^2] = dt$.

Under these assumptions the evolution of the conditional state is governed by the stochastic master equation [3]

$$d\varrho_c = -i[\hat{H}_0, \varrho_c]\, dt + \kappa \mathcal{D}[\hat{a}]\varrho_c\, dt + \sqrt{\eta\kappa}\mathcal{H}[\hat{a}]\varrho_c\, dw_t, \tag{4}$$

where

$$\mathcal{D}[\hat{a}]\varrho_c = \hat{a}\varrho_c\hat{a}^\dagger - \frac{\hat{a}^\dagger\hat{a}\varrho_c + \varrho_c\hat{a}^\dagger\hat{a}}{2}, \tag{5}$$

$$\mathcal{H}[\hat{a}]\varrho_c = \hat{a}\varrho_c + \varrho_c\hat{a}^\dagger - \langle \hat{a} + \hat{a}^\dagger\rangle_c \varrho_c. \tag{6}$$

Notice that the sequence of measures $\tilde{y}_t = \{dy_s\}_{s=0}^t$, $0 \le s \le t$ determines the trajectory followed by the conditional state ϱ_c up to time t, and that the value of ω determines the conditional joint probability density $p(\tilde{y}_t|\omega)$.

In particular, the stochastic master equation (4) for the conditional state ϱ_c is completely equivalent to the equations for its first-moment vector $\bar{\mathbf{r}}_c$ and covariance matrix $\boldsymbol{\sigma}_c$ [59,66,67]:

$$d\bar{\mathbf{r}}_c = A\bar{\mathbf{r}}_c\, dt + (E - \boldsymbol{\sigma}_c B)\frac{\mathbf{dw}_t}{\sqrt{2}}, \tag{7}$$

$$\frac{d\boldsymbol{\sigma}_c}{dt} = A\boldsymbol{\sigma}_c + \boldsymbol{\sigma}_c A^{\mathrm{T}} + D - (E - \boldsymbol{\sigma}_c B)(E - \boldsymbol{\sigma}_c B)^{\mathrm{T}}. \tag{8}$$

The continuous measurement outcome (3) can be written in vectorial form as $\mathbf{dy}_t = -\sqrt{2}B^{\mathrm{T}}\bar{\mathbf{r}}_c\, dt + \mathbf{dw}_t$, with $\mathbf{dw}_t$ the vector of uncorrelated Wiener increments also entering Eq. (7). We refer the reader to Appendix B for details on the matrices entering Eqs. (7) and (8).

It is important to note here that the dynamics determined by the above equations is stable, (i.e., leads to a steady state) if and only if the Hurwitz condition $\mathrm{Re}(\mathrm{eigs}\, A) < 0$ is satisfied; that is, if the real part of the eigenvalues of the drift matrix A is strictly smaller than zero. In our case it corresponds to the inequality $\chi < |\kappa/2|$, and we always assume that this condition is fulfilled.

As we pointed out before, the Hamiltonian in Eq. (1) is able to generate squeezing. If we focus on the unmonitored (unconditional) dynamics (i.e., for $\eta = 0$), the maximum squeezing at the steady state is obtained in the case of $\omega = 0$, leading to a steady-state variance of the $\hat{q}$ quadrature $\langle \Delta\hat{q}^2\rangle_{\mathrm{unc}} = \kappa/(2\kappa + 4\chi)$ that is indeed below the vacuum limit $\langle \Delta\hat{q}^2\rangle_0 = 1/2$ for $0 < \chi < \kappa/2$ (for negative values of χ, one would obtain squeezing along the $\hat{p}$ quadrature). We observe, in particular, that the squeezing increases as we approach instability and that for $\chi \approx \kappa/2$ the well-known limit of $\xi_{\mathrm{dB}} = 3$ dB of squeezing is saturated [68,69]. The Riccati equation (8) can be solved analytically in this case, and its solution shows that continuous monitoring allows us to greatly enhance the squeezing generation for the conditional states ϱ_c. In particular, at the steady state and for $\eta = 1$, a variance $\langle \Delta\hat{q}^2\rangle_c = (\kappa - 2\chi)/2\kappa$ is obtained, thus approaching infinite squeezing near criticality.

For $\omega \neq 0$, no analytical solution is available, but we find by numerical means that a smaller, but still beyond the 3-dB limit, amount of squeezing can be obtained at the steady state; in this case, moreover, the maximum value of the squeezing corresponds, in general, to quadratures different from $\hat{q}$ and $\hat{p}$.

III. FREQUENCY ESTIMATION, SQUEEZING, AND FEEDBACK OPTIMIZATION

Our goal is to devise a protocol able to estimate the frequency parameter ω with high precision. In particular we compare the performance achieved by our proposal with the performances of different strategies that are detailed later herein. In all these strategies, information on the unknown parameter is obtained form two sources: the continuous measurement outcome y_t and a final strong measurement on the corresponding conditional states ϱ_c.

The observation above is made rigorous by our observing the form of the corresponding quantum Cramér-Rao bound that applies in this scenario. As is customary in the context of frequency estimation, we consider the total time of the experiment T, divided into M single runs of duration $t = T/M$, as a fixed resource [70]. Under this assumption, one proves that the precision $\delta\omega$ of any possible unbiased estimator is lower bounded as [28]

$$\delta\omega\, \sqrt{T} \ge \frac{1}{\sqrt{\mathcal{Q}_{\mathrm{eff}}/t}}, \tag{9}$$

where we have defined the *effective* quantum Fisher information (QFI) [12,28]

$$\mathcal{Q}_{\mathrm{eff}} = \mathcal{F}_{\mathrm{hom}} + \bar{\mathcal{Q}}_c, \tag{10}$$

and where we observe that in this scenario the quantity to be optimized is $\mathcal{Q}_{\mathrm{eff}}/t$. The first term, defined as $\mathcal{F}_{\mathrm{hom}} = \mathcal{F}[p(\tilde{y}_t|\omega)]$, corresponds to the classical Fisher information of the conditional probability of observing a trajectory given the value of the parameter ω, and thus to the information obtainable via the continuous homodyne detection [8,11]. The second term, which we define as $\bar{\mathcal{Q}}_c = \mathbb{E}_{\mathrm{traj}}[\mathcal{Q}[\varrho_c]]$, is the average of the QFI $\mathcal{Q}[\varrho_c]$ of the different conditional states generated by the measurement: it thus quantifies the average information obtainable via a final measurement on the different trajectory-dependent ϱ_c (we have introduced the notation $\mathbb{E}_{\mathrm{traj}}[\cdot]$ to denote the average over the conditional distribution $p(\tilde{y}_t|\omega)$ of the different trajectories defined by the stream of measurement outcomes $\tilde{y}_t$). Both quantities can be numerically obtained via the evolution of the first and second moments of the conditional states $\mathbf{r}_c$ and $\boldsymbol{\sigma}_c$, via their derivatives with respect to the parameter (i.e., $\partial_\omega \mathbf{r}_c$ and $\partial_\omega \boldsymbol{\sigma}_c$) and by

performing a Monte Carlo average of the trajectories (see Appendix A for more details). According to Eq. (9), $\mathcal{Q}_{\rm eff}/t$ will thus act as our figure of merit to assess the different estimation protocols that we discuss in the next sections.

The feedback strategy we consider later exploits the information obtained from the continuous measurement output dy_t to perform a unitary feedback operation [3] via the Hamiltonian $\hat{H}_{\rm fb} = \omega_{\rm fb}(t)\hat{a}^\dagger\hat{a}$; that is, by changing either the laser or the cavity resonance frequency via the (possibly time-dependent) parameter $\omega_{\rm fb}(t)$.

To better understand the motivation for a machine learning approach for the optimization of such a feedback strategy, it is expedient to discuss the peculiar features of the estimation problem we are considering. As we discussed in Sec. II, via the Hamiltonian $\hat{H}_0$ in Eq. (1), by fixing $\omega = 0$ and by assuming a positive coupling $\chi > 0$, we know that unconditional squeezing is generated for the quadrature $\hat{q}$, and that a maximum of 3 dB can be obtained at the steady state near instability (i.e., for $\chi \approx \kappa/2$) [68,69]. However, if we also include a continuous homodyne detection, such as the one described by the stochastic master equation (4), the squeezing of the conditional states can be greatly enhanced, going well beyond the 3-dB limit.

The continuous monitoring has, however, also another effect on the conditional state; that is, it gives a stochastic nonzero value for the first moments as described in Eq. (7). Squeezing and nonzero first moments are the relevant figures of merit for the estimation problem we are considering. Squeezing by itself is typically the most important resource for frequency estimation (or analogously for phase estimation [61,62]). However, its interplay with nonzero first moments may play a crucial role in determining the estimation precision. A heuristic representation of this fact is given in Fig. 1: we observe that squeezing could further enhance the estimation if the squeezed quadrature is orthogonal to the direction of the first-moment vector $\bar{\mathbf{r}}_c$ in phase space. Remarkably, if $|\bar{\mathbf{r}}_c|$ is large enough, squeezing for the quadrature parallel to the direction of $\bar{\mathbf{r}}_c$ is going to be detrimental for the estimation of ω. There is a nontrivial trade-off between the amount of squeezing and $|\bar{\mathbf{r}}_c|$, as for small enough $|\bar{\mathbf{r}}_c|$, squeezing along the *wrong* direction is still going to be a useful resource for estimation. We thus expect that the RL agent will be able to optimize such a nontrivial problem by devising feedback strategies able not only to generate large squeezing but also to generate nonzero first moments, and more importantly, to adjust their relative directions in phase space. In general, a final non-Gaussian measurement on the conditional states may be needed to saturate the corresponding quantum Cramér-Rao bound. However, as demonstrated in phase-estimation protocols with Gaussian states [61,71] and as we describe in Appendix E for our results, a final homodyne detection will extract, in general, a fair amount of the maximum amount of information, being nearly optimal for pure Gaussian states.

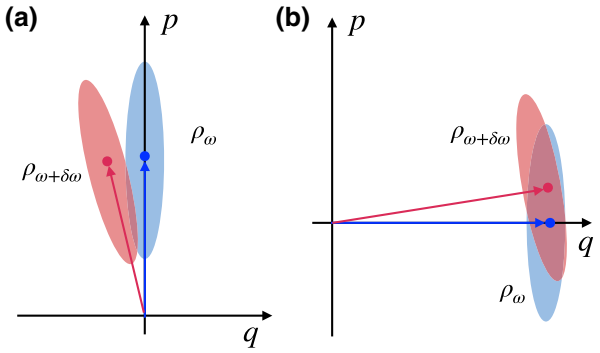

FIG. 1. A heuristic representation in phase space of frequency (phase) estimation via a displaced squeezed state. The estimation precision on the parameter ω can be understood both qualitatively and quantitatively [72] in terms of the distinguishability of quantum states ϱ_ω and $\varrho_{\omega+\delta\omega}$ differing by an infinitesimal value $\delta\omega$ of the parameter. In (a) we observe that if the state is displaced along the $\hat{p}$ axis of the phase space and is squeezed along $\hat{q}$, the two states are highly distinguishable, and thus ω can be measured with high precision. However, as one can see in (b), the quantum states become very indistinguishable in the opposite case (i.e., for both displacement and squeezing along $\hat{q}$). Clearly the situation is much worse for large values of $|\bar{\mathbf{r}}_c|$, while for small first moments, the benefit of squeezing may still yield a high estimation precision.

IV. APPLYING REINFORCEMENT LEARNING

As we mentioned in Sec. I, RL deals with a reward-based learning paradigm. An agent learns how to achieve a certain goal by performing actions on an environment, obtaining complete or partial information on its state, and a reward, specifically designed for the goal.

In this framework the agent is trained over a number of simulations with finite duration called "episodes." At each time step dt, we give the agent access to all the possible information on the conditional state and on its dependence on the parameter ω; that is, by considering as observations the set of parameters $\mathbb{O} = \left(\bar{\mathbf{r}}_c, \boldsymbol{\sigma}_c, \partial_\omega\bar{\mathbf{r}}_c, \partial_\omega\boldsymbol{\sigma}_c, \mathbf{dy}_t\right)$. All these quantities can be updated at each time according to Eqs. (7), (8), (B15), and (B16) once the continuous measurement result $\mathbf{dy}_t$ is obtained. The agent then performs an action on the environment, which in our case consists directly in the choice of a real value for the feedback parameter $\omega_{\rm fb}$.

One of the most important steps in defining a RL problem is to identify the correct reward function. As discussed in the previous section, we assess our feedback strategies via the effective QFI per unit time $\mathcal{Q}_{\rm eff}/t$. We first observe that the Fisher information corresponding to the continuous homodyne detection can be written as (see Appendix B for more details)

$$\mathcal{F}_{\rm hom} = \mathbb{E}_{\rm traj}\left[2\int dt(\partial_\omega\bar{\mathbf{r}}_c)^{\rm T}BB^{\rm T}(\partial_\omega\bar{\mathbf{r}}_c)\right]. \qquad (11)$$

As a consequence we may write

$$\frac{\mathcal{Q}_{\rm eff}}{t} = \mathbb{E}_{\rm traj}\left[\mathcal{R}\right], \tag{12}$$

where we have defined a (positive) trajectory-dependent quantity

$$\mathcal{R} = \frac{2\int dt(\partial_\omega \bar{\mathbf{r}}_c)^{\rm T} BB^{\rm T}(\partial_\omega \bar{\mathbf{r}}_c) + \mathcal{Q}[\varrho_c]}{t}. \tag{13}$$

This observation allows us to state that the maximization of $\mathcal{Q}_{\rm eff}/t$ corresponds to the *trajectory-wise* maximization of $\mathcal{R}$ that will thus act as our reward function (we recall the fact that, as described in Appendix B, $\mathcal{Q}[\varrho_c]$ can be easily evaluated from the properties of the Gaussian conditional state ϱ_c).

We use here the algorithm Proximal Policy Optimization (PPO) [73], a state-of-the-art actor-critic algorithm where the agent is a neural network optimizing both its evaluation of the future reward (critic) and its reward maximization strategy (actor). We exploited the implementation of PPO available in the package STABLE-BASELINES [74]. For this algorithm the strategy, also called a "policy," is a stochastic one, meaning that the action of the agent is extracted from a Gaussian distribution.

The agent we train is a neural network with a feed-forward and fully connected architecture, composed of an input layer of the size of the observations connected to two distinct 64×64 networks (one for the actor and one for the critic). The network is trained with use of a gradient-descent method with linearly decreasing learning rate starting from a value of 2.5×10^{-4} and an entropy coefficient of 0.001 and discount factor $\gamma = 0.99$. At every step of training the loss function is evaluated on batches of 512 elements given by the experience of four parallel workers over a time horizon of 128 time steps. The total number of time steps included in the training is 30×10^6, composed of consecutive simulations (episodes) with a finite duration of 10^5 steps. At the beginning of each episode, the initial condition for the system is set randomly. More specifically, both components of $\mathbf{r}_c$ are set to be extracted from a uniform distribution on the interval $[-3, 3]$, while the number of initial thermal excitations in the system is extracted from a uniform distribution on the interval $[0, 5]$.

V. RESULTS

In the following we fix the unknown, but fixed, frequency, the squeezing rate, and the efficiency of the homodyne measurement, respectively, to $\omega = 0.1\kappa$, $\chi = 0.49\kappa$, and $\eta = 0.9$, with κ the cavity loss rate [see Eq. (3)]. Our simulations show, however, that the agent is able to devise optimized feedback strategies in different regimes; in Appendix D we exemplify such flexibility of the proposed method by showing the results obtained for different values of the monitoring efficiency η.

We denote our figure of merit [i.e., the effective QFI in Eq. (10)] as $\mathcal{Q}_{\rm eff}^{\rm (RL)}$. We compare and contrast the results obtained by means of RL with those obtained by two benchmark strategies: one where *no control* is applied, quantified by the figure of merit $\mathcal{Q}_{\rm eff}^{(0)}$, and the strategy where, thanks to some *a priori* information on the parameter ω (a typical assumption in the context of local quantum estimation theory [72]), a deterministic value of the control frequency is fixed as $\omega_{\rm fb} = -\omega$. In the latter case the control is deterministic and thus it corresponds not to a feedback but rather to an *open-loop* (OL) control strategy, yielding the largest amount of conditional squeezing along the quadrature $\hat{q}$. The continuous monitoring, on the other hand, will yield a nonzero (but typically small) stochastic contribution on the $\hat{q}$ axis of phase space. As a consequence, the directions of squeezing and first moments will not be optimized. We denote the figure of merit for this open-loop control strategy as $\mathcal{Q}_{\rm eff}^{\rm (OL)}$.

The main result of this work is presented in Fig. 2. We consider as the initial state a thermal state with $n_{\rm th} = 5$ thermal excitations and a first-moment vector $\bar{\mathbf{r}}_c = (0, 0)$. We show that, apart from an initial transient time where $\mathcal{Q}_{\rm eff}^{\rm (RL)} \lesssim \mathcal{Q}_{\rm eff}^{\rm (OL)}$, the feedback protocol yields a much larger effective QFI than the benchmark strategies considered. In particular, by looking at the behavior of the two terms entering Eq. (10), we can make two main observations: (i) as regards the average QFIs of the conditional states, which in Fig. 2 correspond to the difference between the curves with the same colors, one finds $\bar{\mathcal{Q}}_c^{(0)} < \bar{\mathcal{Q}}_c^{\rm (RL)} < \bar{\mathcal{Q}}_c^{\rm (OL)}$ (i.e., the feedback protocol is able to generate conditional states that are, on average, more sensitive than the ones generated without feedback, but much less sensitive than the ones generated via the open-loop control protocol); (ii) the enhancement in the estimation is thus obtained mainly thanks to the information contained in the continuous measurement outcomes (the monitoring Fisher information $\mathcal{F}_{\rm hom}^{\rm (RL)}$ greatly overcomes the values of the same figure of merit for the two other protocols). While for the open-loop control protocol $\bar{\mathcal{Q}}_c^{\rm (OL)}$ saturates at a given value once $\boldsymbol{\sigma}_c$ has reached its deterministic steady state, the RL agent seems able to keep $\mathcal{F}_{\rm hom}^{\rm (RL)}$ increasing steadily in time, yielding a large enhancement in the long-time limit.

Moreover we observe that the strategy devised by the agent clearly gives the best result, as it yields the largest values for the figure of merit $\mathcal{Q}_{\rm eff}/t$ appearing in Eq. (9). Our results hint also at the fact that in this case the optimization is obtained in the long-time limit, where the whole information is basically completely contained in the continuous homodyne measurement outcomes and the strong measurement on the conditional states is almost irrelevant (we, however, refer the reader to Appendix E for a discussion on the effectiveness of homodyne detection as a final strong measurement for the three strategies considered).

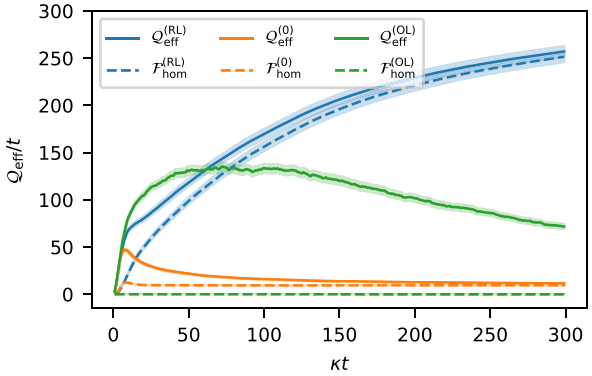

FIG. 2. Comparison of the performance of the feedback strategy devised by the neural network with the performances of the two benchmark strategies described in the main text as a function of time. Solid lines correspond to the effective QFI ($\mathcal{Q}_{\rm eff}^{(\rm RL)}$, $\mathcal{Q}_{\rm eff}^{(0)}$, and $\mathcal{Q}_{\rm eff}^{(\rm OL)}$) divided by time, while dashed lines correspond to the continuous-monitoring classical Fisher information ($\mathcal{F}_{\rm hom}^{(\rm RL)}$, $\mathcal{F}_{\rm hom}^{(0)}$, and $\mathcal{F}_{\rm hom}^{(\rm OL)}$) divided by time. The average QFI of the conditional states ($\bar{\mathcal{Q}}_c^{(\rm RL)}$, $\bar{\mathcal{Q}}_c^{(0)}$, and $\bar{\mathcal{Q}}_c^{(\rm OL)}$) can be derived as the difference between the two curves above. The results are obtained by simulating $N = 5000$ trajectories with time step $dt = 0.001/\kappa$, by fixing the other parameters as $\omega = 0.1\kappa$, $\chi = 0.49\kappa$, and $\eta = 0.9$, and by considering as an initial state a thermal state with $n_{\rm th} = 5$ and initial first-moment vector $\bar{\mathbf{r}}_c(0) = (0, 0)$.

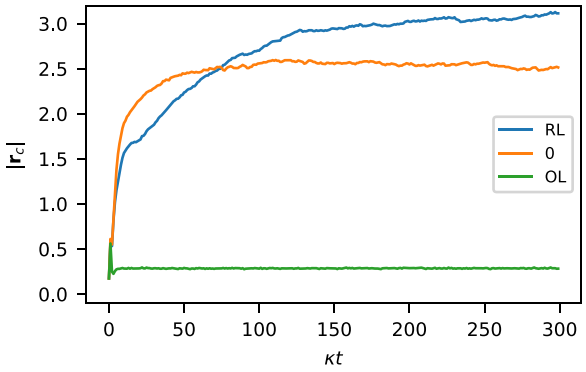

FIG. 3. $\mathbb{E}_{\rm traj}[|\bar{\mathbf{r}}_c|]$ as a function of time for the three protocols considered (from top to bottom, RL feedback, no control, and open-loop control). The results are obtained by simulating $N = 5000$ trajectories with time step $dt = 0.001/\kappa$, by fixing the other parameters as $\omega = 0.1$, $\chi = 0.49\kappa$, and $\eta = 0.9$, and by considering as an initial state a thermal state with $n_{\rm th} = 5$ and initial first-moment vector $\bar{\mathbf{r}}_c(0) = (0, 0)$.

To better understand these results, it is useful to look at the evolution of single trajectories and thus at the properties of the conditional states. As mentioned before, the achievable estimation precision will depend on the amount of squeezing generated during the dynamics and on its interplay with the first-moment vector $\bar{\mathbf{r}}_c$. In Fig. 3 we compare the values of the magnitude of the first moments averaged over the trajectories $\mathbb{E}[|\bar{\mathbf{r}}_c|]$ for the three protocols. We observe that the RL agent yields the largest values of $\mathbb{E}[|\bar{\mathbf{r}}_c|]$, while the OL control protocol yields almost negligible first moments.

We stressed before how squeezing is the main resource for this kind of estimation. In this respect, we know that the maximum amount of squeezing is generated deterministically in the OL control protocol, yielding at the steady state $\xi^{(\rm OL)} \approx 6.05$ dB of squeezing for the values we consider in these simulations. However, we discussed before that this squeezing is always *parallel* to the corresponding vector $\bar{\mathbf{r}}_c^{(\rm OL)}$; despite this fact and because the first moments are close to zero, this protocol still yields large values of $\mathcal{Q}[\varrho_c^{(\rm OL)}]$, as we indeed observe in Fig. 2.

If we now focus on the squeezing along the quadrature *perpendicular* to $\bar{\mathbf{r}}_c$ and thus possibly enhancing the contribution due to nonzero first moments in phase space, we find nontrivial and definitely interesting results as shown in Fig. 4. In this figure we plot the histograms corresponding to the probability density of squeezing perpendicular to the conditional first-moment vector $\bar{\mathbf{r}}_c$ for the no-control protocol and for the RL agent-based feedback protocol for different times. As regards the protocol without control, we know that at the steady state one obtains a deterministic squeezing of $\xi^{(0)} \approx 5.25$ dB, and as a consequence the squeezing along the quadrature perpendicular to the stochastically varying $\bar{\mathbf{r}}_c$ is bounded by this value. This behavior is indeed confirmed by looking at the orange histograms. If we now finally focus on the results corresponding to the RL agent–based feedback (blue histograms), we can clearly observe how it is indeed able also to generate a large fraction of trajectories with squeezing perpendicular to $\bar{\mathbf{r}}_c$ not only well beyond the maximum value obtainable without control $\xi^{(0)}$ but also near the limit $\xi^{(\rm OL)}$ achieved by the open-loop control discussed before and that we recall here is, however, always parallel to the first-moment vector $\bar{\mathbf{r}}_c$. In particular, we observe not only that the RL strategy is able to generate conditional states with the maximum squeezing achievable and with the most useful direction but also that the mode of this perpendicular squeezing distribution quickly saturates toward this limit $\xi^{(\rm OL)}$. Our results thus suggest how the portion of trajectories characterized by large first moments and large perpendicular squeezing is responsible for the enhancement in the frequency estimation precision. We refer the reader to Appendixes C and D for some extra results that we obtained by considering different values of the coupling constant χ and of the monitoring efficiency η, and that further confirm our intuitions. For example, when one considers smaller values of χ, and as a consequence a smaller amount of squeezing generated, all the strategies considered yield as expected smaller values of the effective QFI. Similarly, we show how for smaller values of η, the effect on the squeezing generation is slightly reduced and that

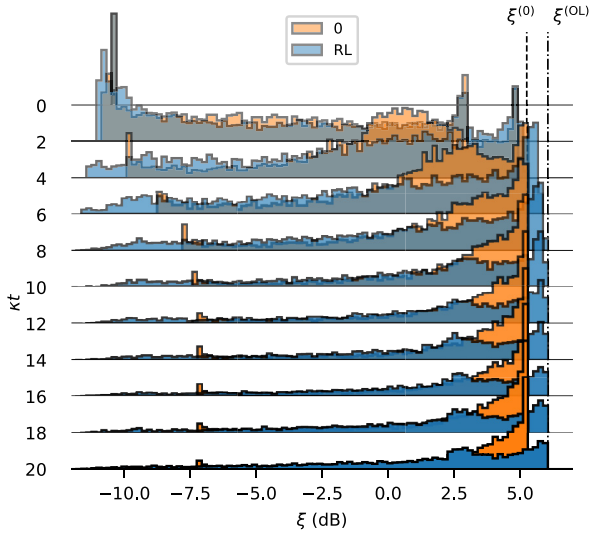

FIG. 4. Histograms of the probability density of the squeezing ξ (expressed in decibels) perpendicular to the conditional first-moment vector $\bar{\mathbf{r}}_c$ for different times and for both the no-control protocol (orange histograms) and the RL agent feedback protocol (blue histograms). The two vertical lines correspond to the two bounds on the amount of squeezing for the no-control protocol $\xi^{(0)}$ and for the open-loop control protocol $\xi^{(\mathrm{OL})}$. The results are obtained by simulating $N = 5000$ trajectories with time step $dt = 0.001/\kappa$, by fixing the other parameters as $\omega = 0.1$, $\chi = 0.49\kappa$, and $\eta = 0.9$, and by considering as an initial state a thermal state with $n_{\mathrm{th}} = 5$ and initial first-moment vector $\bar{\mathbf{r}}_c(0) = (0, 0)$.

the main contribution to the enhancement is given by the first-moment vector amplitude $|\bar{\mathbf{r}}_c|$.

It is also interesting to observe the behavior of the feedback parameter ω_{fb} as a function of time, both for a sample trajectory and averaged over the different trajectories. In Fig. 5 we find that the average value seems to converge to a value near $\mathbb{E}[\omega_{\mathrm{fb}}] \approx -\omega$; that is, the one implemented in the open-loop control and yielding the maximum squeezing. However, at the trajectory level, the fluctuations of ω_{fb} are evident and are thus crucial to increase $|\bar{\mathbf{r}}_c|$ and to optimize both the squeezing magnitude and more importantly its direction.

Plainly speaking, we can conclude that the feedback devised by the neural network is able to optimize the nontrivial interplay between first moments and squeezing and indeed to generate a significant amount of trajectories with a larger amount of *perpendicular squeezing*. These trajectories are thus responsible for the enhancement in the estimation precision observed in Fig. 2. Our results show also that this feature is much more relevant for the homodyne Fisher information $\mathcal{F}_{\mathrm{hom}}$, which is indeed responsible for the enhancement yielded by the feedback strategy. A hint in this direction is already given by Eq. (11) for $\mathcal{F}_{\mathrm{hom}}$, which depends directly on the vector $\partial_\omega \bar{\mathbf{r}}_c$ [however, the evolution of $\partial_\omega \bar{\mathbf{r}}_c$ in Eq. (B15) depends also on $\boldsymbol{\sigma}_c$ and thus on the squeezing properties of the conditional states].

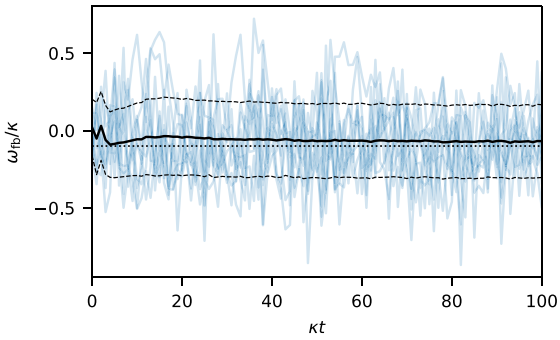

FIG. 5. Feedback parameter ω_{fb} obtained via the RL agent as a function of time, both for a few sample trajectories (corresponding to blue lines with different shades of blue) and averaged over 5000 trajectories (black line). The dashed black lines show the standard deviation. The average of the RL is close to the value $-\omega$, yielding the maximum squeezing (dotted black line). The other physical parameters are set as in the previous figures.

VI. CONCLUSIONS

In this work we show how a RL algorithm is able to optimize a feedback strategy able to attain a high precision in frequency estimation. We understand the results in terms of the optimization of the interplay between the amplitude and the squeezing generated by the protocol. This kind of optimization is highly nontrivial: a simple strategy trying to optimize this kind of feature at each time t cannot be devised because of the stochasticity of the subsequent evolution.

As future work we aim to optimize the neural network so as to be able to reduce the observations $\mathcal{O}$ needed. In particular, we will look at strategies able to exploit just the real-time measurement output $\mathbf{dy}_t$, and thus corresponding to Markovian feedback [32,33]

We have witnessed a great experimental improvement in the implementation of field-programmable gate array–based real-time state-based feedback, as shown recently in the context of the cooling of mechanical oscillators [42–44]. Once a neural network has been trained, its real-time interrogation is not much more computationally costly than what has been done in the cited experiments. We are thus confident that feedback strategies previously trained via RL algorithms can be efficiently implemented in the near future for quantum metrology purposes as we have described, or for more general quantum technological tasks.

ACKNOWLEDGMENTS

We thank F. Albarelli and M. Paris for helpful discussions. M.A.C.R. acknowledges financial support from the Academy of Finland via the Centre of Excellence program (Project No. 336810). M.G.G. and D.T. acknowledge support from the Sviluppo UniMi 2018 initiative. The

computer resources of the Finnish IT Center for Science (CSC) and the FGCI project (Finland) are acknowledged.

APPENDIX A: QUANTUM METROLOGY WITH A CONTINUOUSLY MONITORED QUANTUM SYSTEM

We start by giving a basic introduction to quantum estimation theory. Let us consider a quantum statistical model; that is, a family of quantum states ϱ_ω parametrized by a parameter ω that we want to estimate. We now suppose to perform M measurements, corresponding to a certain positive operator-valued measure $\{\Pi_x\}$, on the quantum state, and thus collect a set of measurement outcomes $\{x_j\}_{j=1}^{M}$. One can prove that the precision of any unbiased estimator $\tilde{\omega}$, which is a map from the measurement outcomes $\{x_j\}$ to the range of parameters taken by ω, is lower bounded according to the Cramér-Rao bound:

$$\delta\omega \geq \frac{1}{\sqrt{M\,\mathcal{F}[p(x|\omega)]}}, \tag{A1}$$

where we have introduced the classical Fisher information

$$\mathcal{F}[p(x|\omega)] = \sum_x \frac{(\partial_\omega p(x|\omega))^2}{p(x|\omega)}, \tag{A2}$$

$$= \mathbb{E}_{p(x|\omega)}\left[\left(\frac{\partial_\omega p(x|\omega)}{p(x|\omega)}\right)^2\right], \tag{A3}$$

and we have denoted with $p(x|\omega) = \mathrm{Tr}[\varrho_\omega \Pi_x]$ the probability of obtaining the outcome x from the measurement. One can further perform optimization over all the possible measurements (positive operator-valued measures) $\{\Pi_x\}$ that one can perform on the quantum state ϱ_ω, obtaining the quantum Cramér-Rao bound

$$\delta\omega \geq \frac{1}{\sqrt{M\,\mathcal{F}[p(x|\omega)]}} \geq \frac{1}{\sqrt{M\,\mathcal{Q}[\varrho_\omega]}}, \tag{A4}$$

where we have introduced the QFI

$$\mathcal{Q}[\varrho_\omega] = \mathrm{Tr}[\varrho_\omega L_\omega^2], \tag{A5}$$

written in terms of the symmetric logarithmic derivative defined via the Lyapunov equation

$$\frac{\partial \varrho_\omega}{\partial \omega} = \frac{L_\omega \varrho_\omega + \varrho_\omega L_\omega}{2}. \tag{A6}$$

Several alternative formulas for the QFI can be derived, based on the diagonalization of the state ϱ_ω or based on the fidelity between states characterized by parameter ω differing by an infinitesimal value [72].

If we want to estimate a parameter in a continuously monitored quantum system, at each run of the experiment we obtain a continuous measurement output (e.g., in the case of continuous homodyne detection) $\tilde{y}_t$ with a certain probability distribution $p_{\mathrm{hom}} = p(\tilde{y}_t|\omega)$ and corresponding to a particular trajectory for the quantum conditional state of the system ϱ_c. In this framework one proves that the bound on the estimation precision can be written as [12]

$$\delta\omega \geq \frac{1}{\sqrt{M\left(\mathcal{F}_{\mathrm{hom}} + \mathbb{E}_{\mathrm{traj}}\left[\mathcal{Q}[\varrho_c]\right]\right)}}. \tag{A7}$$

The relevant figure of merit is thus the effective QFI

$$\mathcal{Q}_{\mathrm{eff}} = \mathcal{F}_{\mathrm{hom}} + \mathbb{E}_{\mathrm{traj}}\left[\mathcal{Q}[\varrho_c]\right], \tag{A8}$$

corresponding to the sum of the Fisher information quantifying the information obtainable from the continuous homodyne results plus the average of the quantum Fisher information of the conditional states, quantifying the information obtainable from a final measurement on ϱ_c.

APPENDIX B: GAUSSIAN CONDITIONAL DYNAMICS AND NUMERICAL EVALUATION OF THE EFFECTIVE QFI

We briefly review here how to treat the evolution of continuously monitored quantum Gaussian states by following the approach in Refs. [59,67] and how to evaluate the different figures of merit relevant for our purposes.

A Gaussian quantum state ϱ of a continuous-variable quantum system is completely identified by its first-moment vector $\bar{\mathbf{r}} = \mathrm{Tr}[\varrho\hat{\mathbf{r}}]$ and its covariance matrix $\boldsymbol{\sigma} = \mathrm{Tr}[\varrho_c\{\hat{\mathbf{r}} - \bar{\mathbf{r}}, (\hat{\mathbf{r}} - \bar{\mathbf{r}})^{\mathrm{T}}\}]$, where $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\} = \hat{\mathbf{a}}\hat{\mathbf{b}}^{\mathrm{T}} + (\hat{\mathbf{b}}\hat{\mathbf{a}}^{\mathrm{T}})^{\mathrm{T}}$. We recall that with this definition the covariance matrix for a single-mode system reads

$$\boldsymbol{\sigma} = 2\begin{pmatrix} \langle\Delta\hat{q}^2\rangle & \langle\Delta\hat{q}\hat{p}\rangle \\ \langle\Delta\hat{q}\hat{p}\rangle & \langle\Delta\hat{p}^2\rangle \end{pmatrix}, \tag{B1}$$

where

$$\langle\Delta\hat{A}\hat{B}\rangle = \mathrm{Tr}\left[\varrho\left(\frac{\hat{A}\hat{B} + \hat{B}\hat{A}}{2}\right)\right] - \mathrm{Tr}[\varrho\hat{A}]\mathrm{Tr}[\varrho\hat{B}]. \tag{B2}$$

The covariance matrix thus directly contains all the squeezing properties of the quantum state ϱ.

As mentioned in the main text, the dynamics induced by the stochastic master equation (4) preserves the Gaussian character of the quantum state, and the corresponding

evolution is described by the equations

$$d\bar{\mathbf{r}}_c = A\bar{\mathbf{r}}_c\, dt + (E - \boldsymbol{\sigma}_c B)\frac{\mathbf{dw}_t}{\sqrt{2}}, \tag{B3}$$

$$\frac{d\boldsymbol{\sigma}_c}{dt} = A\boldsymbol{\sigma}_c + \boldsymbol{\sigma}_c A^{\mathrm{T}} + D - (E - \boldsymbol{\sigma}_c B)(E - \boldsymbol{\sigma}_c B)^{\mathrm{T}}, \tag{B4}$$

while the continuous homodyne outcome is written in vectorial form as

$$\mathbf{dy}_t = -\sqrt{2}B^{\mathrm{T}}\bar{\mathbf{r}}_c\, dt + \mathbf{dw}_t. \tag{B5}$$

The matrices entering these equations can be derived by different approaches [66,67], and for the physical setup we are interested in read

$$A = \begin{pmatrix} -(\chi + \kappa/2) & \omega \\ -\omega & \chi - \kappa/2 \end{pmatrix}, \tag{B6}$$

$$D = \kappa\, \mathbb{1}_2, \tag{B7}$$

$$B = E = \begin{pmatrix} -\sqrt{\eta\kappa} & 0 \\ 0 & 0 \end{pmatrix}. \tag{B8}$$

We observe that the matrices B and E are singular. The second component of the Wiener increment $\mathbf{dw}_t$ in Eq. (B5), therefore, does not play any role at all, whereas the first component is determined by the homodyne detection output.

The solution of the Riccati equation for the covariance matrix (8) can, in general, be obtained numerically. However, an analytical solution can be obtained for the steady-state covariance matrix for $\omega = 0$ and by assuming stable dynamics (i.e., $\chi < \kappa/2$), leading to

$$\boldsymbol{\sigma}_c^{\mathrm{ss}}(\eta) = \begin{pmatrix} \frac{\kappa(2\eta-1)-2\chi+\sqrt{\kappa^2-4\kappa\chi(2\eta-1)+4\chi^2}}{2\eta\kappa} & 0 \\ 0 & \frac{\kappa}{\kappa-2\chi} \end{pmatrix}. \tag{B9}$$

Two opposite regimes can be observed here. By taking the limit for the efficiency $\eta \to 0$, and assuming a stable dynamics, we obtain the solution for the unconditional (unmonitored) dynamics:

$$\boldsymbol{\sigma}_{\mathrm{unc}}^{\mathrm{ss}} = \begin{pmatrix} \frac{\kappa}{\kappa+2\chi} & 0 \\ 0 & \frac{\kappa}{\kappa-2\chi} \end{pmatrix}. \tag{B10}$$

We thus find that for $0 < \chi < \kappa/2$, the Hamiltonian squeezes the $\hat{q}$ quadrature, with a maximum amount of 3 dB of squeezing near instability; that is, for $\chi \to \kappa/2$ [68,69]. In the opposite case of perfect monitoring (i.e., for $\eta = 1$), we find

$$\boldsymbol{\sigma}_c^{\mathrm{ss}} = \begin{pmatrix} \frac{\kappa-2\chi}{\kappa} & 0 \\ 0 & \frac{\kappa}{\kappa-2\chi} \end{pmatrix}, \tag{B11}$$

which in turn, for $0 < \chi < \kappa/2$, corresponds to even smaller variances of the $\hat{q}$ quadrature, and in principle infinite squeezing near instability. For $\omega \neq 0$, we numerically find that a lower amount of squeezing can be generated and that the most squeezed quadrature depends on the value of ω itself.

As described in Appendix A, the performance of the metrological protocol is quantified by the effective QFI defined in Eq. (A8). The quantum states being Gaussian, also this figure of merit can be derived from the information contained in first and second moments. In particular, the homodyne classical Fisher information can be evaluated as [37]

$$\mathcal{F}_{\mathrm{hom}} = \mathbb{E}_{\mathrm{traj}}\left[2\int dt\, (\partial_\omega \bar{\mathbf{r}}_c)^{\mathrm{T}} B B^{\mathrm{T}} (\partial_\omega \bar{\mathbf{r}}_c)\right], \tag{B12}$$

while the QFI of the (Gaussian) conditional state is obtained via the formula [75]

$$\mathcal{Q}[\varrho_c] = \frac{\mathrm{Tr}\left[\left(\boldsymbol{\sigma}_c^{-1}(\partial_\omega \boldsymbol{\sigma}_c)\right)^2\right]}{2(1+\mu^2)} + \frac{2(\partial_\omega \mu)}{1-\mu^4} + 2(\partial_\omega \bar{\mathbf{r}}_c)^{\mathrm{T}} \boldsymbol{\sigma}_c^{-1} (\partial_\omega \bar{\mathbf{r}}_c), \tag{B13}$$

with

$$\mu = \mathrm{Tr}[\varrho_c^2] = 1/\sqrt{\det[\boldsymbol{\sigma}_c]} \tag{B14}$$

denoting the purity of the conditional quantum state. We thus also need the evolution of the derivatives of first and second moments with respect to the parameter ω, which can be numerically integrated via the equations [37]

$$d(\partial_\omega \bar{\mathbf{r}}_c) = (\partial_\omega A)\bar{\mathbf{r}}_c\, dt + A(\partial_\omega \bar{\mathbf{r}}_c)\, dt - \frac{(\partial_\omega \boldsymbol{\sigma}_c) B\, d\mathbf{w}_t}{\sqrt{2}} + (E - \boldsymbol{\sigma}_c B) B^{\mathrm{T}} (\partial_\omega \bar{\mathbf{r}}_c)\, dt, \tag{B15}$$

$$\frac{d(\partial_\omega \boldsymbol{\sigma}_c)}{dt} = (\partial_\omega A)\boldsymbol{\sigma}_c + \boldsymbol{\sigma}_c (\partial_\omega A)^{\mathrm{T}} + A(\partial_\omega \boldsymbol{\sigma}_c) + (\partial_\omega \boldsymbol{\sigma}_c) A^{\mathrm{T}} + (\partial_\omega \boldsymbol{\sigma}_c) B (E - \boldsymbol{\sigma}_c B)^{\mathrm{T}} + (E - \boldsymbol{\sigma}_c B) B^{\mathrm{T}} (\partial_\omega \boldsymbol{\sigma}_c). \tag{B16}$$

At each time t, also the purity μ and its derivative $\partial_\omega \mu$ can be directly obtained from $\boldsymbol{\sigma}_c$ and $(\partial_\omega \boldsymbol{\sigma}_c)$ via Eq. (B14).

These quantities are also exploited as *observations* for the neural network that optimizes the feedback strategy. To

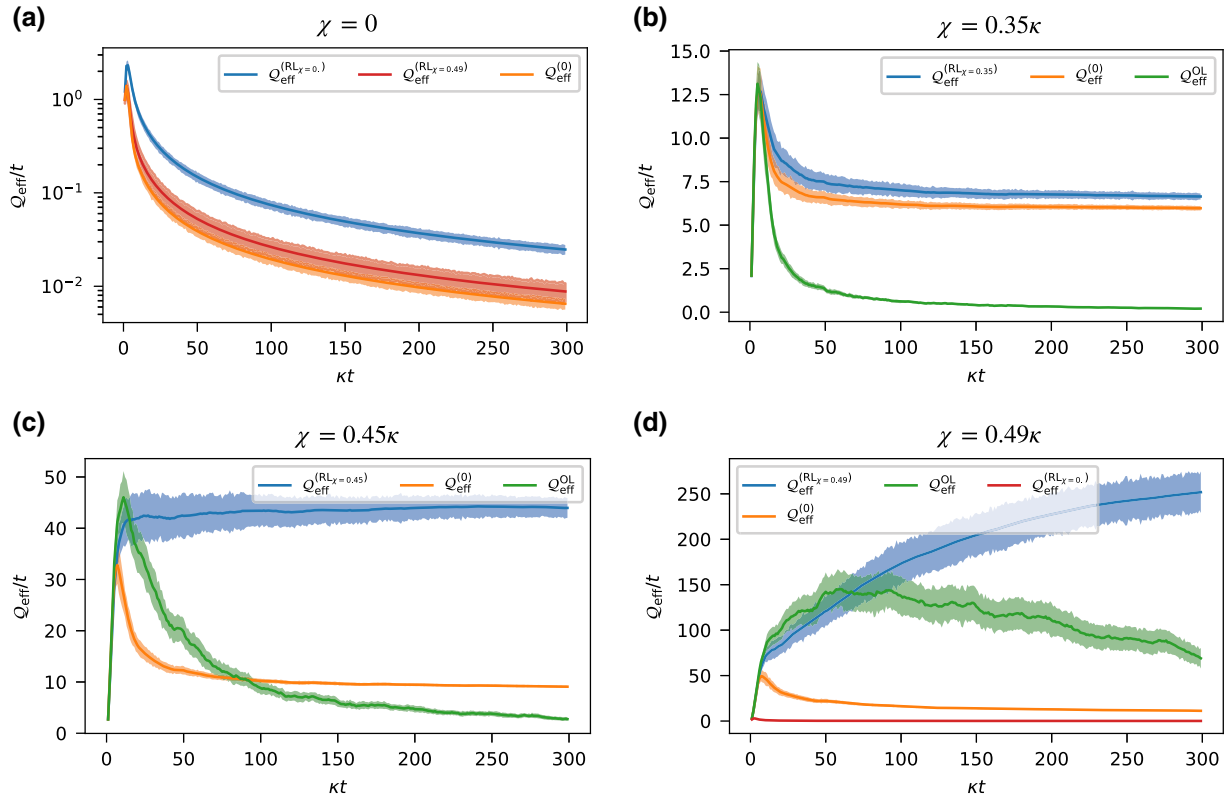

FIG. 6. Performance of the control strategies as a function of time, quantified by the effective quantum Fisher information $\mathcal{Q}_{\rm eff}$ divided by time, and for different values of χ: (a) $\chi = 0$; (b) $\chi = 0.35\kappa$; (c) $\chi = 0.45\kappa$; (d) $\chi = 0.49\kappa$. Notice that in (a) the scale of the y axis is logarithmic and that the curve corresponding to the open-loop control strategy $\mathcal{Q}_{\rm eff}^{(\rm OL)}$ is not reported as the corresponding values are almost negligible. Furthermore in (a),(d) an extra red curve, corresponding to the agent trained at, respectively, $\chi = 0.49$ and $\chi = 0$, has been added. The results are obtained by simulating $N = 1000$ trajectories with time step $dt = 0.001/\kappa$, by fixing the other parameters as $\omega = 0.1\kappa$ and $\eta = 0.9$, and by considering as an initial state a thermal state with $n_{\rm th} = 5$ and initial first-moment vector $\bar{\mathbf{r}}_c(0) = (0, 0)$.

train our agent and to assess the performance of the different protocols, we thus numerically simulate different trajectories of the quantum states via Eqs. (B3), (B4), (B15), and (B16), and we perform the numerical integration and the numerical average in Eqs. (B12) and (A8).

APPENDIX C: EFFECT OF THE HAMILTONIAN COUPLING CONSTANT χ

Here we discuss the role of the Hamiltonian coupling constant χ in the learning of the strategy by the agent and in its performances. As highlighted in the main text, χ is

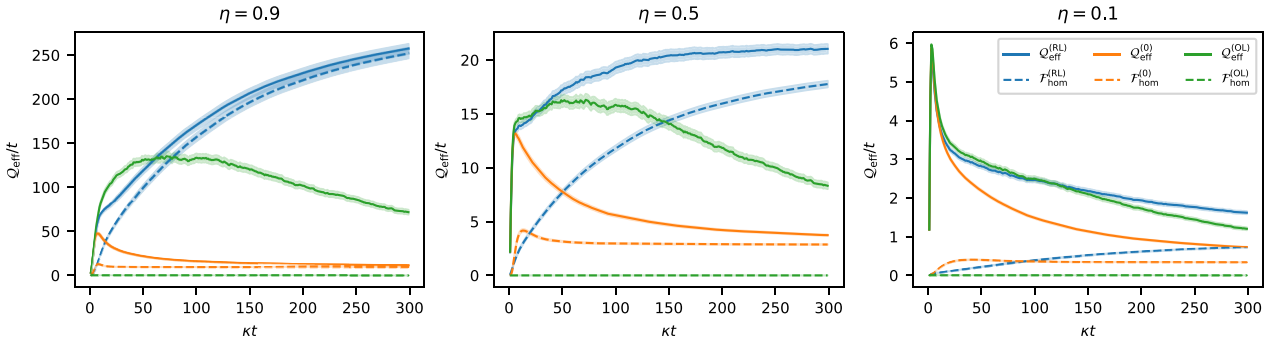

FIG. 7. Performance of the control strategies as a function of time, quantified by the Fisher information divided by time, and for different values of η. All the other parameters are fixed as in Fig. 2.

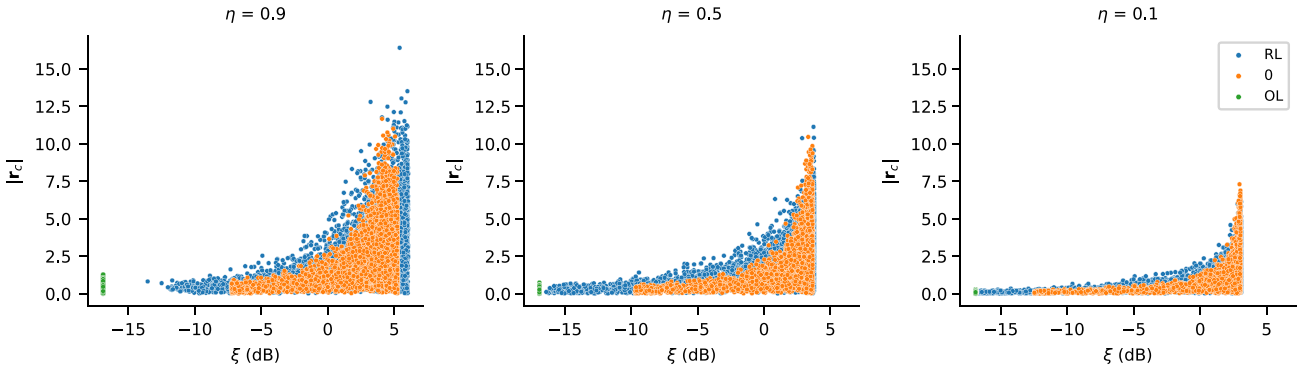

FIG. 8. Scatter plot of the squeezing (expressed in decibels) perpendicular to the conditional first-moment vector $\bar{\mathbf{r}}_c$ and the absolute value of the first-moment vector $|\bar{\mathbf{r}}_c|$ for the three control strategies at time $\kappa t = 180$ and for different values of η.

directly responsible for the generation of squeezing in the conditional states. Squeezing can be observed if and only if χ is larger than zero, and in particular near criticality (i.e., for $\chi \approx \kappa/2$, the amount of squeezing generated is close to infinity in the case of perfect monitoring).

In Fig. 6 we compare our control strategy with the benchmark strategies by plotting our figure of merit $\mathcal{Q}_{\rm eff}/t$, obtained by agents trained with different values of χ (0, 0.35κ, 0.45κ, and 0.49κ), as a function of time. We observe how the agent allows us to reach values of the QFI larger than those obtained by means of the other strategies, in particular in the long-time limit. Other relevant observations can be drawn from these plots: (i) In general, we find that, for all strategies, larger values of χ yield larger values of the QFI, thus highlighting, once again, the importance of the squeezing generated during the dynamics. (ii) For values of χ that are *large enough* (e.g., for $\chi = 0.45\kappa$ and $\chi = 0.49\kappa$), the agent is able to devise a strategy such that the maximum of $\mathcal{Q}_{\rm eff}/t$ is observed in the long-time limit; for smaller values of χ (e.g., $\chi = 0.35\kappa$) one observes a maximum at short times, while in the long-time limit the ratio between the effective QFI and time tends to a smaller stationary value. (iii) We also find that for $\chi = 0.45\kappa$ the maximum of $\mathcal{Q}_{\rm eff}/t$ obtained for the open-loop control strategy is compatible with the value obtained in the long-time limit via the agent strategy. We stress, however, that while for the open-loop strategy one would need to stop the dynamics at a very specific time, by using the RL strategy, one has a wide available time window, in which the dynamics has reached a steady-state behavior.

To better describe the properties of the strategies devised by the neural network, we have added an extra curve in Figs. 6(a) and 6(d): in Fig. 6(a) we have added the values of the QFI corresponding to the agent trained at $\chi = 0.49\kappa$ applied to the case $\chi = 0$, while in Fig. 6(d) we have added the performance of the agent trained at $\chi = 0$ applied to the case $\chi = 0.49\kappa$. In the first scenario we observe that the agent trained at $\chi = 0.49\kappa$ is still able to beat the benchmark strategies and its performance is only slightly poorer than that of the properly trained agent, In the second scenario the situation is completely inverted: the agent trained at $\chi = 0$ yields very low values of the QFI. Our interpretation of this result is as follows: when the agent is trained near criticality, it learns how to optimize both first moments and squeezing, and thus performs well also when squeezing is absent. On the other hand, at $\chi = 0.49\kappa$ (i.e., when squeezing plays a major role in the estimation protocol) the agent trained in the no-squeezing scenario completely fails in enhancing the estimation precision.

APPENDIX D: EFFECT OF CONTINUOUS MONITORING OF EFFICIENCY

Here we present some results that we obtained for different values of the monitoring efficiency η (0.9, 0.5, and 0.1). In Fig. 7 we report the behavior of the different Fisher information divided by time as in Fig. 2. We observe how the enhancement is still clearly observed for $\eta = 0.5$ and that even for very small monitoring efficiency ($\eta = 0.1$) the agent feedback is able to yield a larger estimation precision compared with the other strategies, in particular in the long-time limit.

We also report in Fig. 8 scatter plots of the perpendicular squeezing defined in the main text and the absolute value of the first-moment vector $|\bar{\mathbf{r}}_c|$ for a fixed time $\kappa t = 180$. We observe as remarked also for Fig. 4 that with respect to the no-control strategy, for $\eta = 0.9$ the agent is able to prepare trajectories with larger perpendicular squeezing. However, in this plot we also observe that the agent is in general able to prepare conditional states that, for a fixed amount of perpendicular squeezing, yields larger first moments, thus leading to an enhanced estimation. By reducing the monitoring efficiency, we find that the first effect (trajectories with larger perpendicular squeezing) is basically lost also for $\eta = 0.5$, while the second effect (i.e., larger first moments at fixed squeezing) is still obtained

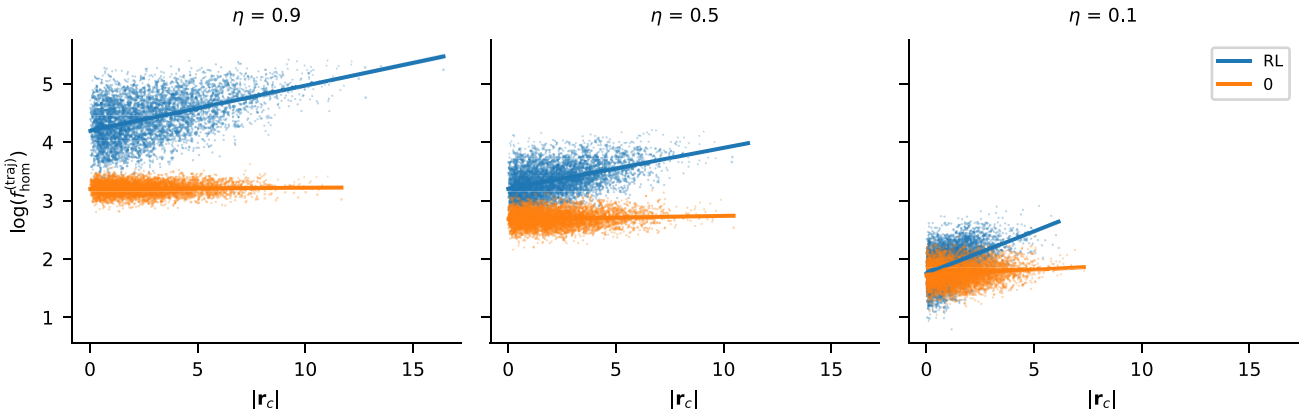

FIG. 9. Scatter plot between the absolute value of the first-moment vector $|\bar{\mathbf{r}}_c|$ and the trajectory contribution to the homodyne Fisher information $\log f_{\rm hom}^{(\rm traj)}$ at time $\kappa t = 180$ and for different values of η. Only the points corresponding to the agent strategy and the no-control strategy are plotted, as $f_{\rm hom}^{(\rm traj)} = 0$ for the OL control strategy.

and thus it can be considered as solely responsible for the greater estimation precision.

The relevance of the first-moment vector is also highlighted in Fig. 9, corresponding to the scatter plot of $|\bar{\mathbf{r}}_c|$ and $\log f_{\rm hom}^{(\rm traj)}$ for the different trajectories, where we have introduced the quantity

$$f_{\rm hom}^{(\rm traj)} = 2\int dt\, (\partial_\omega \bar{\mathbf{r}}_c)^{\rm T} BB^{\rm T} (\partial_\omega \bar{\mathbf{r}}_c), \tag{D1}$$

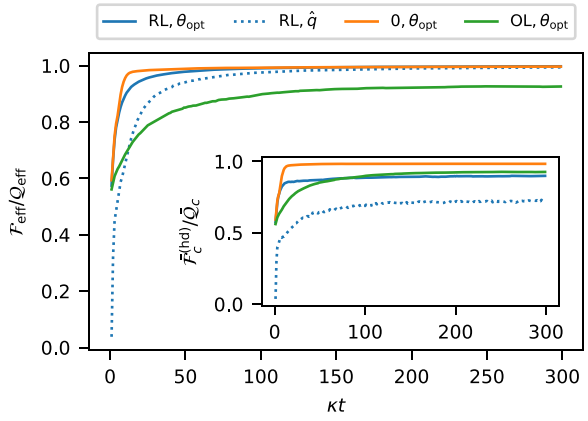

FIG. 10. Ratio between the effective Fisher information $\mathcal{F}_{\rm eff}$ for final homodyne detection and the effective QFI of the conditional state $\bar{\mathcal{Q}}_{\rm eff}$ as a function of κt for the three different strategies (the inset shows the ratio between the average Fisher information of the strong homodyne measurement $\bar{\mathcal{F}}_c^{(\rm hd)}$ and the average QFI $\bar{\mathcal{Q}}_c$.) The solid lines show the Fisher information $\bar{\mathcal{F}}_{\rm hd,opt}$ optimized over θ, while the dotted line shows the Fisher information $\bar{\mathcal{F}}_{\rm hd,\hat{q}}$ obtained with a homodyne on the quadrature $\hat{q}$ for the RL strategy. In all cases, while not being optimal, homodyne detection allows one to extract a significant fraction of the total available information on the frequency ω (parameter values are fixed as in Fig. 2).

corresponding to the contribution of each trajectory to the homodyne Fisher information $\mathcal{F}_{\rm hom} = \mathbb{E}[f_{\rm hom}^{(\rm traj)}]$ [see Eq. (11)]. In Fig. 9 we observe how the two quantities seem to be correlated for the agent strategy, while they seem uncorrelated for the no-control strategy (we recall that for the OL control strategy $f_{\rm hom}^{(\rm traj)} = 0$ for each trajectory), highlighting once again the mechanism behind the strategy devised by the agent.

APPENDIX E: EFFECTIVENESS OF HOMODYNE DETECTION AS A FINAL STRONG MEASUREMENT

Here we discuss the effectiveness of homodyne detection as a final strong measurement in our protocol (i.e., we calculate the Fisher information of a final homodyne measurement and we compare it with the QFI $\bar{\mathcal{Q}}_c$). In general, a non-Gaussian measurement may be needed to saturate the quantum Cramér-Rao bound (i.e., to obtain classical Fisher information equal to the QFI); however, we expect that homodyne detection will extract a fair amount of the maximum information achievable, being nearly optimal for pure Gaussian states (see Refs. [61,71] for an extensive study of phase estimation with Gaussian states).

A projective Gaussian measurement can be modeled by the covariance matrix of a squeezed vacuum state, with squeezing parameter s, and a phase rotation of angle θ:

$$\boldsymbol{\sigma}_m(z,\theta) = R(\theta)\,\mathrm{diag}(z, 1/z)\,R(\theta)^{\rm T}, \tag{E1}$$

where $z = \exp 2s$ and $R(\theta)$ is a rotation matrix. In the limit $z \to 0$, we have a homodyne measurement: when $\theta = 0$ ($\theta = \pi/2$), the quadrature $\hat{q}$ ($\hat{p}$) is measured.

The measurement outcome probability of such a measurement on a state with first moments $\mathbf{r}_c$ and covariance matrix $\boldsymbol{\sigma}_c$ is a two-dimensional Gaussian distribution $\mathcal{N}(\mathbf{r}_c, \Sigma)$, where $\Sigma = (\boldsymbol{\sigma}_c + \boldsymbol{\sigma}_m)/2$, for which the Fisher

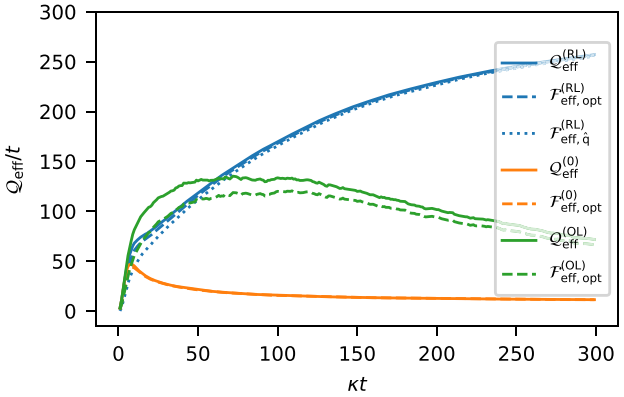

FIG. 11. $\mathcal{Q}_{\rm eff}/t$ (solid lines) and $\mathcal{F}_{\rm eff}/t$ for homodyne detection (dashed lines) as a function of κt for the three different strategies under consideration. Use of homodyne detection as a final strong measurement does not affect the precision in the estimation significantly for the RL strategy, given that the biggest contribution to $\mathcal{Q}_{\rm eff}$ comes from the continuous monitoring. The dotted blue line shows $\mathcal{F}_{\rm eff}/t$ for homodyne detection on the $\hat{q}$ quadrature for the RL strategy. Even without optimizing the angle of the final homodyne measurement, the RL strategy allows one to exceed the performances of the benchmark strategies (parameter values are fixed as in Fig. 2).

information reads

$$\mathcal{F}[\mathcal{N}(\mathbf{r}_c,\Sigma)|\omega] = (\partial_\omega \boldsymbol{r}_c)^{\mathsf{T}}\Sigma^{-1}\partial_\omega \boldsymbol{r}_c + \frac{1}{2}\mathrm{Tr}[(\Sigma^{-1}\partial_\omega\Sigma)^2]. \tag{E2}$$

For each trajectory, we maximize the Fisher information for a final homodyne measurement over $\theta \in [-\pi/2, \pi/2]$, $\mathcal{F}_c^{(\rm hd)}$. We then calculate the average $\bar{\mathcal{F}}_c^{(\rm hd)}$ and compare it with $\bar{\mathcal{Q}}_c$. The resulting ratio, $\bar{\mathcal{F}}_c^{(\rm hd)}/\bar{\mathcal{Q}}_c$, is shown in Fig. 10 for the three different strategies. As can be seen, the optimized homodyne detection, although not ideal, allows the extraction of a significant amount of information from the quantum state. Finding the optimal angle θ in real time should be possible by means, for example, of a field-programmable gate array (see the discussion on the feedback real-time implementation in Sec. VI). The dotted lines in the inset in Fig. 10 show that even if a fixed, and suitably *a priori* chosen, angle θ is used, it is still possible to extract a significant fraction of the QFI for the state (we set $\theta = 0$ for the RL strategy, corresponding to measuring $\hat{q}$, as suggested by a direct inspection of the distribution of the optimal θ values for the different trajectories). The overall effect on the effective Fisher information when homodyne detection is used as a strong final measurement is shown in Fig. 10: in the strategy devised by the RL agent, where the contribution of the final strong measurement to $\mathcal{Q}_{\rm eff}$ is small, the ratio between the effective Fisher information $\mathcal{F}_{\rm eff}$ corresponding to a final homodyne detection (either optimized or for $\theta = 0$) and the effective QFI $\mathcal{Q}_{\rm eff}$ tends rapidly to 1 as the monitoring time κt is increased. We also notice that the ratio is reasonably high also for the benchmark strategies: the more significant effect can be seen in the OL control scenario, particularly at short times, because $\mathcal{Q}_{\rm eff} = \bar{\mathcal{Q}}_c$.

We finally compare the performance of a final homodyne detection for the three strategies in Fig. 11, where we observe that the enhancement obtained via the RL strategy is still maintained when a final homodyne detection is performed.

[1] V. Giovannetti, S. Lloyd, and L. Maccone, Advances in quantum metrology, Nat. Photonics 5, 222 (2011).

[2] S. Pirandola, B. R. Bardhan, T. Gehring, C. Weedbrook, and S. Lloyd, Advances in photonic quantum sensing, Nat. Photonics **12**, 724 (2018).

[3] H. M. Wiseman and G. J. Milburn, *Quantum Measurement and Control* (Cambridge University Press, New York, 2010).

[4] K. Jacobs and D. A. Steck, A straightforward introduction to continuous quantum measurement, Contemp. Phys. 47, 279 (2006).

[5] M. Guţă, B. Janssens, and J. Kahn, Optimal estimation of qubit states with continuous time measurements, Commun. Math. Phys. 277, 127 (2007).

[6] M. Tsang, H. M. Wiseman, and C. M. Caves, Fundamental Quantum Limit to Waveform Estimation, Phys. Rev. Lett. **106**, 090401 (2011).

[7] M. Tsang, Quantum metrology with open dynamical systems, New J. Phys. 15, 73005 (2013).

[8] S. Gammelmark and K. Mølmer, Bayesian parameter inference from continuously monitored quantum systems, Phys. Rev. A 87, 032115 (2013).

[9] S. Gammelmark and K. Mølmer, Fisher Information and the Quantum Cramér-Rao Sensitivity Limit of Continuous Measurements, Phys. Rev. Lett. **112**, 170401 (2014).

[10] M. Guţă and J. Kiukas, Information geometry and local asymptotic normality for multi-parameter estimation of quantum markov dynamics, J. Mat. Phys. **58**, 052201 (2017).

[11] M. G. Genoni, Cramér-Rao bound for time-continuous measurements in linear Gaussian quantum systems, Phys. Rev. A 95, 012116 (2017).

[12] F. Albarelli, M. A. C. Rossi, M. G. A. Paris, and M. G. Genoni, Ultimate limits for quantum magnetometry via time-continuous measurements, New J. Phys. 19, 123011 (2017).

[13] H. Mabuchi, Dynamical identification of open quantum systems, Quant. Semiclass. Opt. **8**, 1103 (1996).

[14] J. Gambetta and H. M. Wiseman, State and dynamical parameter estimation for open quantum systems, Phys. Rev. A **64**, 042105 (2001).

[15] J. M. Geremia, J. K. Stockton, A. C. Doherty, and H. Mabuchi, Quantum Kalman Filtering and the Heisenberg Limit in Atomic Magnetometry, Phys. Rev. Lett. **91**, 250801 (2003).

[16] K. Mølmer and L. B. Madsen, Estimation of a classical parameter with Gaussian probes: Magnetometry

with collective atomic spins, Phys. Rev. A 70, 052102 (2004).
[17] L. B. Madsen and K. Mølmer, Spin squeezing and precision probing with light and samples of atoms in the Gaussian description, Phys. Rev. A **70**, 052324 (2004).
[18] J. K. Stockton, J. M. Geremia, A. C. Doherty, and H. Mabuchi, Robust quantum parameter estimation: Coherent magnetometry with feedback, Phys. Rev. A 69, 032109 (2004).
[19] M. Tsang, Optimal waveform estimation for classical and quantum systems via time-symmetric smoothing. II. Applications to atomic magnetometry and Hardy's paradox, Phys. Rev. A 81, 013824 (2010).
[20] T. A. Wheatley, D. W. Berry, H. Yonezawa, D. Nakane, H. Arao, D. T. Pope, T. C. Ralph, H. M. Wiseman, A. Furusawa, and E. H. Huntington, Adaptive Optical Phase Estimation Using Time-Symmetric Quantum Smoothing, Phys. Rev. Lett. **104**, 093601 (2010).
[21] H. Yonezawa, D. Nakane, T. A. Wheatley, K. Iwasawa, S. Takeda, H. Arao, K. Ohki, K. Tsumura, D. W. Berry, T. C. Ralph, H. M. Wiseman, E. H. Huntington, and A. Furusawa, Quantum-enhanced optical phase tracking, Science **337**, 1514 (2012).
[22] R. L. Cook, C. A. Riofrío, and I. H. Deutsch, Single-shot quantum state estimation via a continuous measurement in the strong backaction regime, Phys. Rev. A 90, 032113 (2014).
[23] P. Six, P. Campagne-Ibarcq, L. Bretheau, B. Huard, and P. Rouchon, in *2015 54th IEEE Conf. Decis. Control*, Cdc (IEEE, 2015), p. 7742.
[24] A. H. Kiilerich and K. Mølmer, Bayesian parameter estimation by continuous homodyne detection, Phys. Rev. A **94**, 032103 (2016).
[25] L. Cortez, A. Chantasri, L. P. García-Pintos, J. Dressel, and A. N. Jordan, Rapid estimation of drifting parameters in continuously measured quantum systems, Phys. Rev. A 95, 012314 (2017).
[26] J. F. Ralph, S. Maskell, and K. Jacobs, Multiparameter estimation along quantum trajectories with sequential monte carlo methods, Phys. Rev. A 96, 052306 (2017).
[27] J. Atalaya, S. Hacohen-Gourgy, L. S. Martin, I. Siddiqi, and A. N. Korotkov, Correlators in simultaneous measurement of non-commuting qubit observables, npj Quantum Inf. 4, 41 (2018).
[28] F. Albarelli, M. A. C. Rossi, D. Tamascelli, and M. G. Genoni, Restoring heisenberg scaling in noisy quantum metrology by monitoring the environment, Quantum **2**, 110 (2018).
[29] A. Shankar, G. P. Greve, B. Wu, J. K. Thompson, and M. Holland, Continuous Real-Time Tracking of a Quantum Phase below the Standard Quantum Limit, Phys. Rev. Lett. **122**, 233602 (2019).
[30] M. A. C. Rossi, F. Albarelli, D. Tamascelli, and M. G. Genoni, Noisy Quantum Metrology Enhanced by Continuous Nondemolition Measurement, Phys. Rev. Lett. **125**, 200505 (2020).
[31] A. C. Doherty and K. Jacobs, Feedback control of quantum systems using continuous state estimation, Phys. Rev. A **60**, 2700 (1999).
[32] H. M. Wiseman and G. J. Milburn, Quantum Theory of Optical Feedback via Homodyne Detection, Phys. Rev. Lett. **70**, 548 (1993).
[33] H. M. Wiseman and G. J. Milburn, Squeezing via feedback, Phys. Rev. A **49**, 1350 (1994).
[34] L. K. Thomsen, S. Mancini, and H. M. Wiseman, Spin squeezing via quantum feedback, Phys. Rev. A 65, 061801 (2002).
[35] A. Serafini and S. Mancini, Determination of Maximal Gaussian Entanglement Achievable by Feedback-Controlled Dynamics, Phys. Rev. Lett. **104**, 220501 (2010).
[36] A. Szorkovszky, A. C. Doherty, G. I. Harris, and W. P. Bowen, Mechanical Squeezing via Parametric Amplification and Weak Measurement, Phys. Rev. Lett. **107**, 213603 (2011).
[37] M. G. Genoni, S. Mancini, and A. Serafini, Optimal feedback control of linear quantum systems in the presence of thermal noise, Phys. Rev. A **87**, 042333 (2013).
[38] M. G. Genoni, J. Zhang, J. Millen, P. F. Barker, and A. Serafini, Quantum cooling and squeezing of a levitating nanosphere via time-continuous measurements, New J. Phys. **17**, 073019 (2015).
[39] S. G. Hofer and K. Hammerer, Entanglement-enhanced time-continuous quantum control in optomechanics, Phys. Rev. A 91, 033822 (2015).
[40] M. Brunelli, D. Malz, and A. Nunnenkamp, Conditional Dynamics of Optomechanical Two-Tone Backaction-Evading Measurements, Phys. Rev. Lett. **123**, 093602 (2019).
[41] A. Di Giovanni, M. Brunelli, and M. G. Genoni, Unconditional mechanical squeezing via backaction-evading measurements and nonoptimal feedback control, Phys. Rev. A **103**, 022614 (2021).
[42] M. Rossi, D. Mason, J. Chen, Y. Tsaturyan, and A. Schliesser, Measurement-based quantum control of mechanical motion, Nature **563**, 53 (2018).
[43] L. Magrini, P. Rosenzweig, C. Bach, A. Deutschmann-Olek, S. G. Hofer, S. Hong, N. Kiesel, A. Kugi, and M. Aspelmeyer, Real-time optimal quantum control of mechanical motion at room temperature, Nature **595**, 373 (2021).
[44] F. Tebbenjohanns, M. L. Mattana, M. Rossi, M. Frimmer, and L. Novotny, Quantum control of a nanoparticle optically levitated in cryogenic free space, Nature **595**, 378 (2021).
[45] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT press, Cambridge, Massachusetts, 2018).
[46] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, Nature **518**, 529 (2015).
[47] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Mastering the game of go without human knowledge, Nature **550**, 354 (2017).

[48] F. Marquardt, Machine Learning and Quantum Devices, SciPost Phys. Lect. Notes, 292021.
[49] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement Learning with Neural Networks for Quantum Feedback, Phys. Rev. X **8**, 031084 (2018).
[50] S. Mavadia, V. Frey, J. Sastrawan, S. Dona, and M. J. Biercuk, Prediction and real-time compensation of qubit decoherence via machine learning, Nat. Commun. **8**, 14106 (2017).
[51] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, Universal quantum control through deep reinforcement learning, npj Quantum Inf. **5**, 33 (2019).
[52] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Coherent transport of quantum states by deep reinforcement learning, Commun. Phys. **2**, 61 (2019).
[53] J. Brown, P. Sgroi, L. Giannelli, G. S. Paraoanu, E. Paladino, G. Falci, M. Paternostro, and A. Ferraro, Reinforcement learning-enhanced protocols for coherent population-transfer in three-level quantum systems (2021), ArXiv:2109.00973.
[54] L. Moro, M. G. A. Paris, M. Restelli, and E. Prati, Quantum compiling by deep reinforcement learning, Commun. Phys. **4**, 178 (2021).
[55] S. Corli, L. Moro, D. E. Galli, and E. Prati, Solving rubik's cube via quantum mechanics and deep reinforcement learning, J. Phys. A: Math. Theor. **54**, 425302 (2021).
[56] S. Borah, B. Sarma, M. Kewming, G. J. Milburn, and J. Twamley, Measurement Based Feedback Quantum Control With Deep Reinforcement Learning, (2021), ArXiv:2104.11856.
[57] R. Porotti, A. Essig, B. Huard, and F. Marquardt, Deep reinforcement learning for quantum state preparation with weak nonlinear measurements (2021), ArXiv:2107.08816.
[58] E. N. Evans, Z. Wang, A. G. Frim, M. R. DeWeese, and E. A. Theodorou, Stochastic optimization for learning quantum state feedback control, (2021), ArXiv:2111.09896.
[59] A. Serafini, *Quantum Continuous Variables* (CRC Press, Boca Raton, 2017).
[60] R. D. Candia, F. Minganti, K. V. Petrovnin, G. S. Paraoanu, and S. Felicetti, Critical parametric quantum sensing (2021), ArXiv:2107.04503.
[61] A. Monras, Optimal phase measurements with pure gaussian states, Phys. Rev. A **73**, 033821 (2006).
[62] M. G. Genoni, S. Olivares, and M. G. A. Paris, Optical Phase Estimation in the Presence of Phase Diffusion, Phys. Rev. Lett. **106**, 153603 (2011).
[63] M. H. Schleier-Smith, I. D. Leroux, and V. Vuletić, States of an Ensemble of Two-Level Atoms with Reduced Quantum Uncertainty, Phys. Rev. Lett. **104**, 073604 (2010).
[64] I. D. Leroux, M. H. Schleier-Smith, and V. Vuletić, Implementation of Cavity Squeezing of a Collective Atomic Spin, Phys. Rev. Lett. **104**, 073602 (2010).
[65] W. Wasilewski, K. Jensen, H. Krauter, J. J. Renema, M. V. Balabas, and E. S. Polzik, Quantum Noise Limited and Entanglement-Assisted Magnetometry, Phys. Rev. Lett. **104**, 133601 (2010).
[66] H. M. Wiseman and A. C. Doherty, Optimal Unravellings for Feedback Control in Linear Quantum Systems, Phys. Rev. Lett. 94, 070405 (2005).
[67] M. G. Genoni, L. Lami, and A. Serafini, Conditional and unconditional Gaussian quantum dynamics, Contemp. Phys. **57**, 331 (2016).
[68] G. Milburn and D. Walls, Production of squeezed states in a degenerate parametric amplifier, Opt. Commun. **39**, 401 (1981).
[69] M. J. Collett and C. W. Gardiner, Squeezing of intracavity and traveling-wave light fields produced in parametric amplification, Phys. Rev. A **30**, 1386 (1984).
[70] S. F. Huelga, C. Macchiavello, T. Pellizzari, A. K. Ekert, M. B. Plenio, and J. I. Cirac, Improvement of Frequency Standards with Quantum Entanglement, Phys. Rev. Lett. 79, 3865 (1997).
[71] C. Oh, C. Lee, C. Rockstuhl, H. Jeong, J. Kim, H. Nha, and S.-Y. Lee, Optimal gaussian measurements for phase estimation in single-mode gaussian metrology, npj Quantum Inf. **5**, 10 (2019).
[72] M. G. A. Paris, Quantum estimation for Quantum technology, Int. J. Quant. Inf. **07**, 125 (2009).
[73] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms, (2017), arXiv preprint ArXiv:1707.06347.
[74] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, Stable baselines, https://github.com/hill-a/stable-baselines 2018.
[75] O. Pinel, P. Jian, N. Treps, C. Fabre, and D. Braun, Quantum parameter estimation using general single-mode Gaussian states, Phys. Rev. A 88, 040102 (2013).